Minimax Capacity Loss under Sub-Nyquist Universal Sampling

Yuxin Chen, Andrea J. Goldsmith, and Yonina C. Eldar

Abstract

This paper considers the capacity of sub-sampled analog channels when the sampler is designed to operate independent of instantaneous channel realizations. A compound multiband Gaussian channel with unknown subband occupancy is considered, with perfect channel state information available at both the receiver and the transmitter. We restrict our attention to a general class of periodic sub-Nyquist samplers, which subsumes as special cases sampling with periodic modulation and filter banks.

We evaluate the loss due to channel-independent (universal) sub-Nyquist design through a *sampled capacity loss* metric, that is, the gap between the undersampled channel capacity and the Nyquist-rate capacity. We investigate sampling methods that minimize the worst-case (minimax) capacity loss over all channel states. A fundamental lower bound on the minimax capacity loss is first developed, which depends only on the band sparsity ratio and the undersampling factor, modulo a residual term that vanishes at high signal-to-noise ratio. We then quantify the capacity loss under Landau-rate sampling with periodic modulation and low-pass filters, when the Fourier coefficients of the modulation waveforms are randomly generated and independent (resp. i.i.d. Gaussian-distributed), termed independent random sampling (resp. Gaussian sampling). Our results indicate that with exponentially high probability, independent random sampling and Gaussian sampling achieve minimax sampled capacity loss in the Landau-rate and super-Landau-rate regime, respectively. While identifying a deterministic minimax sampling scheme is in general intractable, our results highlight the power of randomized sampling methods, which are optimal in a universal design sense. Similar results and conclusions for a discrete-time sparse vector channel can be delivered as an immediate consequence of our analysis: independent random sensing matrices and i.i.d. Gaussian matrices are respectively minimax in the Landau-rate and super-Landau-rate regime.

Index Terms

Channel capacity, sub-Nyquist sampling, universal sampling, non-asymptotic random matrix, minimaxity, log-determinant, concentration of measure, universality phenomenon, compressed sensing, discrete-time sparse vector channel

Y. Chen is with the Department of Electrical Engineering and the Department of Statistics, Stanford University, Stanford, CA 94305, USA (email: yxchen@stanford.edu). A. J. Goldsmith is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA (email: andrea@ee.stanford.edu). Y. C. Eldar is with the Department of Electrical Engineering, Technion, Israel Institute of Technology Haifa, Israel 32000, and has been a visiting professor at Stanford University (email: yonina@ee.technion.ac.il). The contact author is Y. Chen.
This work was supported in part by the NSF Center for Science of Information, the Interconnect Focus Center of the Semiconductor Research Corporation, and BSF Transformative Science Grant 2010505. It will be presented in part at the IEEE International Symposium on Information Theory (ISIT) 2013. Manuscript date: May 2, 2013.

I. INTRODUCTION

The maximum rate of information that can be conveyed through an analog communication channel largely depends on the sampling technique and rate employed at the receiver end. In wideband communication systems, hardware and cost limitations often preclude sampling at or above the Nyquist rate, which presents a major bottleneck in transferring wideband and energy-efficient receiver design paradigms from theory to practice. Understanding the effects upon capacity of sub-Nyquist sampling is thus crucial in circumventing this bottleneck.

In practice, receiver hardware and, in particular, sampling mechanisms are typically static and hence designed based on a family of possible channel realizations. During operation, the actual channel realization will vary over this class of channels. Since the sampler is typically integrated into the hardware and difficult to change during system operation, it needs to be designed independent of instantaneous channel state information (CSI). This has no effect if the sampling rate employed is commensurate with the maximum bandwidth (or the Nyquist rate) of the channel family. However, at the sub-Nyquist sampling rate regime, the sampler design significantly impacts the information rate achievable over different channel realizations. As was shown in [1], the capacity-maximizing sub-Nyquist sampling mechanism for a given linear time-invariant (LTI) channel depends on specific channel realizations. In time-varying channels, *sampled capacity loss* relative to the Nyquist-rate capacity is necessarily incurred due to channel-independent (universal) sub-Nyquist sampler design. Moreover, it turns out that the capacity-optimizing sampler for a given channel structure might result in very low data rate for other channel realizations.

In this paper, our goal is to explore universal design of a sub-Nyquist sampling system that is robust against the uncertainty and variation of instantaneous channel realizations, based on sampled capacity loss as a metric. In particular, we investigate the fundamental lower limit of sampled capacity loss in some overall sense (as will be detailed as minimax capacity loss in Section II-C), and design a sub-Nyquist sampling system for which the capacity loss can be uniformly controlled and optimized over all possible channel realizations.

A. Related Work

In various scenarios, sampling at or above the Nyquist rate is not necessary for preserving signal information if certain signal structures are appropriately exploited [2], [3]. Take multiband signals for example, that reside within several subbands over a wide spectrum. If the spectral support is known, then the sampling rate necessary for perfect signal reconstruction is the spectral occupancy, termed the *Landau rate* [4]. Such signals admit perfect recovery when sampled at rates approaching the Landau rate, assuming appropriately chosen sampling sets (e.g. [5], [6]). Inspired by recent "compressive sensing" [7]–[9] ideas, spectrum-blind sub-Nyquist samplers have also been developed for multiband signals [10], pulse streams [11], [12], etc. These sampling-theoretic works, however, were not based on capacity as a metric in the sampler design.

On the other hand, the Shannon-Nyquist sampling theorem has frequently been used to investigate analog waveform channels (e.g. [13]–[17]). One key paradigm to determine or bound the channel capacity is converting the continuous-time channel into a set of parallel discrete-time channels, under the premise that sampling, when it is performed at or above the Nyquist rate, preserves information. In addition, the effects upon capacity of oversampling

have been investigated in the presence of quantization [18], [19]. However, none of these works considered the effect of reduced-rate sampling upon capacity. Another recent line of work [20] investigated the tradeoff between sparse coding and subsampling in AWGN channels, but did not consider capacity-achieving input distributions.

Our recent work [1], [21] established a new framework for investigating the capacity of linear time invariant (LTI) Gaussian channels under a broad class of sub-Nyquist sampling strategies, including filter-bank and modulation-bank sampling [10], [22] and, more generally, time-preserving sampling. We showed that periodic sampling or, more simply, sampling with a filter bank, is sufficient to approach maximum capacity among all sampling structures under a given sampling rate constraint, assuming that perfect CSI is available at both the receiver and the transmitter.

Practical communication systems often involve time-varying channels, e.g. wireless slow fading channels [16], [23]. Many of these channels can be modeled as a channel with state (see a detailed survey in [23, Chapter 7]), where the channel variation is captured by a state that may be fixed over a long transmission block, or, more simply, a compound channel [24] whereby the channel realization lies within a collection of possible channels [25]. One class of compound channel models concerns multiband Gaussian channels, whereby the instantaneous frequency support active for transmission resides within several continuous intervals, spread over a wide spectrum. This model naturally arises in several wideband communication systems, including time division multiple access systems and cognitive radio networks, as will be discussed in Section II-A. However, to the best of our knowledge, no prior work has investigated, from a capacity perspective, a universal (channel-independent) sub-Nyquist sampling paradigm that is robust to channel variations in the above channel models.

Finally, we note that the design of optimal sampling / sensing matrices have recently been investigated in discrete-time settings from a information theoretical perspective. In particular, Donoho *et. al.* [26] assert that: random and band-diagonal sampling systems admit perfect signal recovery from an information theoretically minimal number of samples. However, the optimality was not defined based on channel capacity as a metric, but instead based on a fundamental rate-distortion limit [27].

B. Contribution

In this paper, we consider a compound multiband channel, whereby the channel bandwidth W is divided into n continuous subbands and, at each timeframe, only k subbands are active for transmission. We consider the class of periodic sampling systems (i.e. a system that consists of a periodic preprocessor and a recurrent sampling set, detailed in Section II-B) with period n/W and sampling rate $f_{\mathrm{s}} = mW/n$ for some integer m ($m \leq n$). Under this model, we define $\beta := \frac{k}{n}$ as the band sparsity ratio, and $\alpha := \frac{m}{n}$ as the undersampling factor. The sampling mechanism is termed a Landau-rate sampling (resp. super-Landau-rate sampling) system if f_{s} is equal to (resp. greater than) the spectral size of the instantaneous channel support. Our contributions are as follows.

1) We derive, in Theorem 4, a fundamental lower bound on the largest sampled capacity loss (defined in Section II) incurred by any channel-independent sampler, under both Landau-rate and super-Landau-rate sampling. This lower bound depends only on the band sparsity ratio and the undersampling factor, modulo a small residual term that vanishes when SNR and n increase. The bound is derived by observing that at each

frequency within $[0, W/n]$, the exponential sum of the capacity loss over all states s is independent of the sampling system, except for a relatively small residual term that vanishes with SNR.

2) Theorem 5 characterizes the sampled capacity loss under a class of periodic sampling with periodic modulation (of period $\frac{n}{W}$) and low-pass filters with passband $\left[0, \frac{W}{n}\right]$, when the Fourier coefficients of the modulation waveforms are randomly generated and independent (termed *independent random sampling*). We demonstrate that with exponentially high probability, the sampled capacity loss matches the fundamental lower bound of Theorem 4 uniformly across all channel realizations. This implies that independent sampling achieves the minimum worst-case (or *minimax*) capacity loss among all periodic sampling methods with period $\frac{n}{W}$. To be more concrete, an independent random sampling system achieves minimax sampled capacity loss as long as the Fourier coefficients of the modulation waveforms are independently sub-Gaussian [28] distributed with matching moments up to the second order. This universality phenomenon occurs due to sharp concentration of spectral measures of large random matrices [29].
3) For a large portion of the super-Landau-rate regime, we quantify the sampled capacity loss under independent random sampling when the Fourier coefficients of the modulation waveforms are i.i.d. Gaussian-distributed (termed *Gaussian random sampling*), as stated in Theorem 6. With exponentially high probability, Gaussian random sampling achieves minimax capacity loss among all periodic sampling with period $\frac{n}{W}$.
4) Similar results for a discrete-time sparse vector channel with unknown channel support can be delivered as an immediate consequence of our analysis framework. When the number of measurements is equal to the channel support size, independent random sensing matrices are minimax in terms of channel-blind sampler design. When the sample size exceeds the channel support size, Gaussian sensing matrices achieve minimax sampled capacity loss.

C. *Organization*

The remainder of this paper is organized as follows. In Section II, we introduce our system model of compound multiband channels. A metric called sampled capacity loss, and a minimax sampler are defined with respect to sampled channel capacity. We then determine in Section III the minimax capacity loss that is achievable within the class of periodic sampling systems. Specifically, we develop a lower bound on the minimax capacity loss in Section III-A. Its achievability under Landau-rate and super-Landau-rate sampling are treated in Section III-B and Section III-C, respectively. Section IV-A summarizes the key observation and implications from our results. We present in Section V extensions to discrete-time sparse vector channels, and discuss connections with compressed sensing literature. Section VI closes the paper with a short summary of our findings and potential future directions.

D. *Notation*

We define the following two functions: $\log^{\epsilon}(x) := \log(\max(\epsilon, x))$, and $\det^{\epsilon} \boldsymbol{X} := \prod_i \max(\lambda_i(\boldsymbol{X}), \epsilon)$. Denote by $\mathcal{H}(\beta) := -\beta \log \beta - (1-\beta)\log(1-\beta)$ the binary entropy function, and $\mathcal{H}(\{x_1, \cdots, x_n\}) := -\sum_{i=1}^{n} x_i \log x_i$ the more general entropy function. The standard notation $f(n) = O(g(n))$ means there exists a constant c (not

Table I

SUMMARY OF NOTATION AND PARAMETERS

$\mathcal{H}(x)$	binary entropy function, i.e. $\mathcal{H}(x)=-x\log x-(1-x)\log(1-x)$
$h(t)$,$H(f)$	impulse response, and frequency response of the LTI analog channel
$\mathcal{S}_{\eta}(f)$	power spectral density of the noise $\eta(t)$
f_{s}, T_{s}	aggregate sampling rate, and the corresponding sampling interval ($T_{\text{s}}=1/f_{\text{s}}$)
W, W_0	channel bandwidth, size of instantaneous channel support
n,m,k	number of subbands, number of sampling branches, number of subbands being simultaneously active
$\alpha=f_{\text{s}}/W=m/n$, $\beta=k/n$	undersampling factor, sparsity ratio
$\boldsymbol{Q}$, $\boldsymbol{Q}^{\text{w}}=(\boldsymbol{Q}\boldsymbol{Q}^{*})^{-\frac{1}{2}}\boldsymbol{Q}$	sampling matrix, whitened sampling matrix
$L_{\boldsymbol{s}}^{\boldsymbol{Q}}$	capacity loss associated with a sampling matrix $\boldsymbol{Q}$ given state $\boldsymbol{s}$
$\log^{\epsilon}(x)$, $\det^{\epsilon}(\boldsymbol{X})$	$\log^{\epsilon}(x):=\log(\max(\epsilon,x))$, $\det^{\epsilon}(\boldsymbol{X}):=\prod_i\max(\lambda_i(\boldsymbol{X}),\epsilon)$
$\boldsymbol{A}_{i*}$, $\boldsymbol{A}_{*i}$	ith row of $\boldsymbol{A}$, ith column of $\boldsymbol{A}$
$\text{card}(\mathcal{A})$	cardinality of a set $\mathcal{A}$
$[n]$	$[n]:=\{1,2,\cdots,n\}$
$\binom{\mathcal{A}}{k}$,$\binom{[n]}{k}$	set of all k-element subsets of $\mathcal{A}$, set of all k-element subsets of $[n]$
$\mathcal{W}_p(n,\Sigma)$	$p\times p$-dimensional central Wishart distribution with n degrees of freedom and covariance matrix Σ
$\mathbb{Z}$, $\mathbb{R}$	set of integers, set of real numbers

necessarily positive) such that $|f(n)|\leq cg(n)$. For a matrix $\boldsymbol{A}$, we use $\boldsymbol{A}_{i*}$ and $\boldsymbol{A}_{*i}$ to denote the ith row and ith column of $\boldsymbol{A}$, respectively. We let $[n]$ denote the set $\{1,2,\cdots,n\}$. For any set $\mathcal{A}\subseteq[n]$, we denote by $\binom{\mathcal{A}}{k}$ the set of all k-combinations of $\mathcal{A}$. In particular, we write $\binom{[n]}{k}$ for the set of all k-element subsets of $\{1,2,\cdots,n\}$. We also use $\text{card}(\mathcal{A})$ to denote the cardinality of a set $\mathcal{A}$. Let $\boldsymbol{W}$ be a $p\times p$ random matrix that can be expressed as $\boldsymbol{W}=\Sigma_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T$, where $\boldsymbol{Z}_i\sim\mathcal{N}(0,\boldsymbol{\Sigma})$ are jointly independent vectors. Then $\boldsymbol{W}$ is said to have a central Wishart distribution with n degrees of freedom and covariance matrix Σ, denoted as $\boldsymbol{W}\sim\mathcal{W}_p(n,\Sigma)$. Our notation is summarized in Table I.

II. PROBLEM FORMULATION AND PRELIMINARIES

A. Compound Multiband Channel

We consider a compound multiband Gaussian channel. The channel has a total bandwidth W, and is divided into n continuous subbands each of bandwidth W/n. A state $\boldsymbol{s}\in\binom{[n]}{k}$ is generated, which dictates the channel support and realization[1]. Specifically, given a state $\boldsymbol{s}$, the channel is an LTI filter with impulse response $h_{\boldsymbol{s}}(t)$ and frequency response $H_{\boldsymbol{s}}(f)=\int_{-\infty}^{\infty}h_{\boldsymbol{s}}(t)\exp(-j2\pi ft)\text{d}t$. It is assumed throughout that there exists a general

[1]Note that in practice, n is typically a large number. For instance, the number of subcarriers ranges from 128 to 2048 in LTE [30], [31].

function $H(f, \boldsymbol{s})$ such that for every f and $\boldsymbol{s}$, $H_{\boldsymbol{s}}(f)$ can be expressed as

$$H_{\boldsymbol{s}}(f) = H(f, \boldsymbol{s})\mathbf{1}_{\boldsymbol{s}}(f), \quad \text{where} \quad \mathbf{1}_{\boldsymbol{s}}(f) = \begin{cases} 1, & \text{if } f \text{ lies within subbands at indices from } \boldsymbol{s}, \\ 0, & \text{else.} \end{cases}$$

A transmit signal $x(t)$ with a power constraint P is passed through this multiband channel, which yields a channel output

$$r_{\boldsymbol{s}}(t) = h_{\boldsymbol{s}}(t) * x(t) + \eta(t), \tag{1}$$

where $\eta(t)$ is stationary zero-mean Gaussian noise with power spectral density $\mathcal{S}_\eta(f)$. We assume that perfect CSI is available at both the transmitter and the receiver.

The above model subsumes as special cases the following communication scenarios.

- **Time Division Multiple Access Model.** In this setting the channel is shared by a set of different users. At each timeframe, one of the users is selected for transmission. The receiver (e.g. the base station) allocates a subset of subbands to the designated sender over that timeframe.
- **White-Space Cognitive Radio Network.** In a white-space cognitive radio network, cognitive users exploit spectrum holes unoccupied by primary users and utilize them for communications. Since the locations of the spectrum holes change over time, the spectral subbands available to cognitive users is varying over time.

B. Sampled Channel Capacity

We aim to design a sampler that works at rates below the Nyquist rate (i.e. the channel bandwidth W). In particular, we consider the class of periodic sampling systems, which subsume the most widely used sampling mechanisms in practice.

1) Periodic Sampling: The class of periodic sampling systems is defined in [1, Section IV], which we restate as follows.

Definition 1 (**Periodic Sampling**). Consider a sampling system consisting of a preprocessor with an impulse response $q(t, \tau)$ followed by a sampling set $\Lambda = \{t_k \mid k \in \mathbb{Z}\}$. A linear sampling system is said to be periodic with period T_q and sampling rate f_s ($f_s T_q \in \mathbb{Z}$) if for every $t, \tau \in \mathbb{R}$ and every $k \in \mathbb{Z}$, we have

$$q(t, \tau) = q(t + T_q, \tau + T_q); \qquad t_{k + f_s T_q} = t_k + T_q. \tag{2}$$

Consider a periodic sampling system $\mathcal{P}$ with period $T_q = n/W$ and sampling rate $f_{\mathrm{s}} := mW/n$ for some integer m. A special case consists of sampling with a combination of filter banks and periodic modulation with period n/W, as illustrated in Fig. 1(a). Specifically, the sampling system comprises m branches, where at each branch, the channel output is passed through a pre-modulation filter, modulated by a periodic waveform of period T_q, and then filtered with a post-modulation filter followed by uniform sampling at rate f_{s}/m. The channel capacity in a sub-sampled LTI Gaussian channel has been derived in [1, Theorem 5]. As will be shown later, in the high SNR regime, employing water-filling power allocation harvests only marginal capacity gain relative to equal power allocation. For

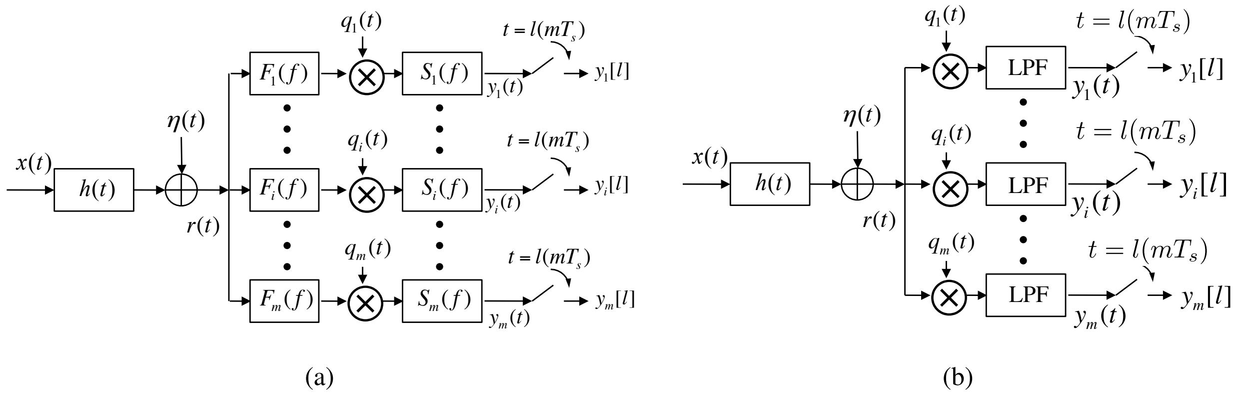


Figure 1. (a) Sampling with modulation and filter banks. The channel output $r(t)$ is passed through m branches, each consisting of a pre-modulation filter, a periodic modulator and a post-modulation filter followed by a uniform sampler with sampling rate f_s/m. (b) Sampling with a modulation bank and low-pass filters. The channel output $r(t)$ is passed through m branches, each consisting of a modulator with modulation waveform $q_i(t)$ and a low-pass filter of pass band $[0, f_s/m]$ followed by a uniform sampler with sampling rate f_s/m.

this reason and for mathematical convenience, we only restate below the sampled channel capacity under uniform power allocation, which suffices to demonstrate the fundamental minimax limit as well as the convergence rate with SNR. Specifically, if we denote by $\boldsymbol{s}_i$ and f_i the ith smallest element in $\boldsymbol{s}$ and the lowest frequency of the ith subband, respectively, and define $\boldsymbol{H}_{\boldsymbol{s}}(f)$ as a $k \times k$ diagonal matrix obeying

$$(\boldsymbol{H}_{\boldsymbol{s}}(f))_{ii} = \frac{|H_{\boldsymbol{s}}\left(f_{\boldsymbol{s}_i} + f\right)|}{\sqrt{\mathcal{S}_\eta\left(f_{\boldsymbol{s}_i} + f\right)}},$$

then the sampled channel capacity, when specialized to our setting, is given as follows.

Theorem 1 (**Sampled Capacity with Equal Power Allocation** [1])**.** *Consider a channel with total bandwidth W and instantaneous band sparsity ratio $\beta := \frac{k}{n}$. Assume perfect CSI at both the transmitter and the receiver, and equal power allocation employed over active subbands. If a periodic sampler $\mathcal{P}$ with period n/W and sampling rate $f_s = \frac{m}{n}W$ is employed, then the sampled channel capacity at a given state $\boldsymbol{s}$ is given by*

$$C_{\boldsymbol{s}}^{\boldsymbol{Q}} = \frac{1}{2}\int_0^{\frac{W}{n}} \log\det\left(\boldsymbol{I}_m + \frac{P}{\beta W}\boldsymbol{Q}^{\mathrm{w}}(f)\boldsymbol{H}_{\boldsymbol{s}}^2(f)\boldsymbol{Q}^{\mathrm{w}*}(f)\right)\mathrm{d}f, \tag{3}$$

where $\boldsymbol{Q}^{\mathrm{w}}(f) := \left(\boldsymbol{Q}(f)\boldsymbol{Q}^*(f)\right)^{-1/2}\boldsymbol{Q}(f)$. Here, $\boldsymbol{Q}(f)$ is an $m \times n$ matrix that only depends on $\mathcal{P}$, and $\boldsymbol{Q}_{\boldsymbol{s}}(f)$ denotes the submatrix consisting of the columns of $\boldsymbol{Q}(f)$ at indices of $\boldsymbol{s}$.

In general, $\boldsymbol{Q}(f)$ is a function that varies with f. Unless otherwise specified, we call $\boldsymbol{Q}(\cdot)$ the *sampling coefficient function* and $\boldsymbol{Q}^{\mathrm{w}}(\cdot)$ the *whitened sampling coefficient function* with respect to the sampling system $\mathcal{P}$. Note that $\boldsymbol{Q}^{\mathrm{w}}(f)\boldsymbol{Q}^{\mathrm{w}*}(f) = \boldsymbol{I}$.

2) Flat Sampling Coefficient Function: A special class of periodic sampling concerns the ones whose $\boldsymbol{Q}(\cdot)$ are flat over $[0, f_s/m]$, in which case we can use an $m \times n$ matrix $\boldsymbol{Q}$ to represent the sampling coefficient function, termed a *sampling coefficient matrix*. This class of sampling systems can be realized through the m-branch sampling system illustrated in Fig. 1(b). In the ith branch, the channel output is modulated by a periodic waveform $q_i(t)$ of

period n/W, passed through a low-pass filter with pass band $[0, f_s/m]$, and then uniformly sampled at rate f_s/m, where the Fourier transform of $q_i(t)$ obeys $\mathcal{F}(q_i(t)) = \sum_{l=1}^{n} \boldsymbol{Q}_{i,l}\delta\left(f - l\frac{W}{n}\right)$. In this paper, a sampling system within this class is said to be *(independent) random sampling* if the entries of $\boldsymbol{Q}$ are randomly generated (and are independent). In addition, a sampling system is termed *Gaussian sampling* if the entries of $\boldsymbol{Q}$ are i.i.d. Gaussian distributed.

It turns out that this simple class of sampling structures is sufficient to achieve overall robustness in terms of sampled capacity loss, provided that the entries of $\boldsymbol{Q}$ are sub-Gaussian with zero mean and unit variance, as will be detailed in Section III.

C. Universal Sampling

As was shown in [1], the optimal sampling mechanism for a given LTI channel with perfect CSI extracts out a frequency set with the highest SNR and hence suppresses aliasing. Such alias-suppressing sampler may achieve very low capacity for some channel realizations. In this paper, we desire a sampler that operates independent of the instantaneous CSI, and our objective is to design a single linear sampling system that achieves to within a minimal gap of the Nyquist-rate capacity across all possible channel realizations. The availability of CSI to the transmitter, the receiver and the sampler is illustrated in Fig. 2.

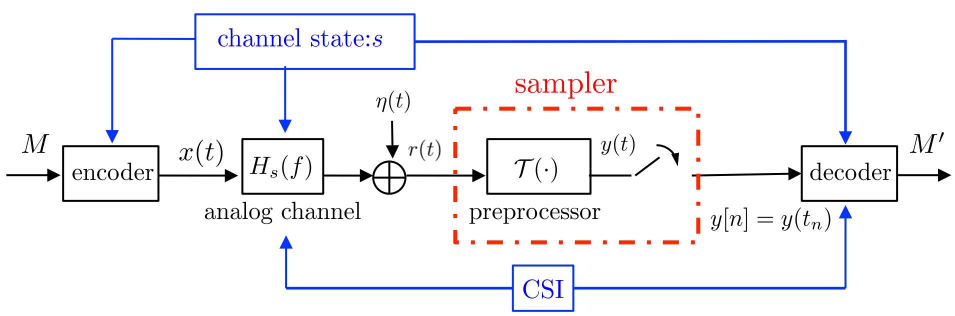


Figure 2. At each timeframe, a state is generated from a finite set $\mathcal{S}$, which dictates the channel realization $H_{\boldsymbol{s}}(f)$. Both the transmitter and the receiver have perfect CSI, while the sampler operates independently of $\boldsymbol{s}$.

1) Sampled Capacity Loss: Universal sub-Nyquist samplers suffer from information rate loss relative to Nyquist-rate capacity. In this subsection, we make formal definitions of this metric.

For any state $\boldsymbol{s}$, when equal power allocation is performed over active subbands, the Nyquist-rate capacity can be written as

$$C_{\boldsymbol{s}}^{P_{\mathrm{eq}}} = \int_0^{W/n} \frac{1}{2} \log\det\left(\boldsymbol{I}_k + \frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}^2(f)\right) \mathrm{d}f, \tag{4}$$

which is a special case of (3). In contrast, if power control at the transmitter side is allowed, then the Nyquist-rate

capacity is given by

$$C_{\boldsymbol{s}}^{\text{opt}} = \int_0^{W/n} \frac{1}{2} \sum_{i=1}^{k} \log^+ \left(\nu \left(\boldsymbol{H}_{\boldsymbol{s}}(f) \right)_{ii}^2 \right) \mathrm{d}f, \tag{5}$$

where ν is determined by the equation

$$P = \int_0^{W/n} \sum_{i=1}^{k} \left(\nu - \frac{1}{\left(\boldsymbol{H}_{\boldsymbol{s}}(f) \right)_{ii}^2} \right)^+ \mathrm{d}f. \tag{6}$$

We can then make formal definitions of sampled capacity loss as follows.

Definition 2 (**Sampled Capacity Loss).** For any sampling coefficient function $\boldsymbol{Q}(\cdot)$ and any given state $\boldsymbol{s}$, we define the sampled capacity loss without power control as

$$L_{\boldsymbol{s}}^{\boldsymbol{Q}} := C_{\boldsymbol{s}}^{P_{\text{eq}}} - C_{\boldsymbol{s}}^{\boldsymbol{Q}},$$

and define the sampled capacity loss with optimal power control as

$$L_{\boldsymbol{s}}^{\boldsymbol{Q},\text{opt}} := C_{\boldsymbol{s}}^{\text{opt}} - C_{\boldsymbol{s}}^{\boldsymbol{Q}}.$$

These metrics quantify the capacity gaps relative to Nyquist-rate capacity due to *universal* (channel-independent) sub-Nyquist sampling design. When sampling is performed at or above the Landau rate (which is equal to $\frac{kW}{n}$ in our case) but below the Nyquist rate, these gaps capture the rate loss due to channel-independent sampling relative to channel-optimized design, both with or without power control.

For notational convenience, for an $m \times n$ matrix $\boldsymbol{M}$, we denote by $L_{\boldsymbol{s}}^{\boldsymbol{M}}$ and $L_{\boldsymbol{s}}^{\boldsymbol{M},\text{opt}}$ the capacity loss with respect to a sampling coefficient function $\boldsymbol{Q}(f) \equiv \boldsymbol{M}$, which is flat across $[0, W/n]$.

2) Minimax sampler: Frequently used in the theory of statistics (e.g. [32]), minimaxity is a metric that seeks to minimize the loss function in some overall sense, defined as follows.

Definition 3 (**Minimax Sampler).** A sampling system associated with a sampling coefficient function $\boldsymbol{Q}^{\text{m}}$, which minimizes the worst-case capacity loss, that is, which satisfies

$$\max_{\boldsymbol{s} \in \binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{Q}^{\text{m}}} = \inf_{\boldsymbol{Q}(\cdot)} \max_{\boldsymbol{s} \in \binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{Q}},$$

is called a minimax sampler with respect to the state alphabet $\binom{[n]}{k}$.

The minimax criteria is of interest for designing a sampler robust to all possible channel states, that is, achieving to within a minimal gap relative to maximum capacity for all channel realizations. It aims to control the rate loss across all states in a uniform manner, as illustrated in Fig. 3. Note that the minimax sampler is in general different from the one that maximizes the lowest capacity among all states (worst-case capacity). While the latter guarantees an optimal worst-case capacity that can be achieved regardless of which channel is realized, it may result in significant capacity loss in many states with large Nyquist-rate capacity, as illustrated in Fig. 3. In contrast,

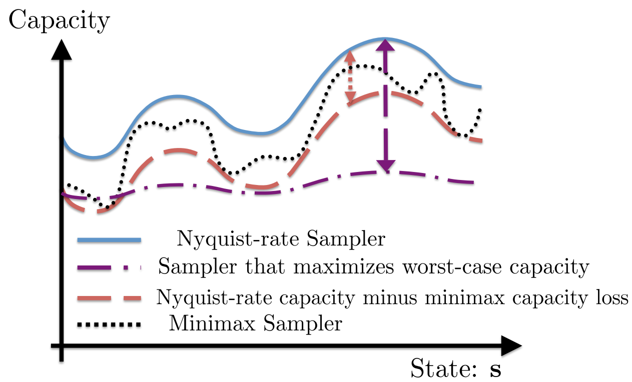


Figure 3. Minimax sampler v.s. the sampler that maximizes worst-case capacity, when sampling is channel-independent and performed below the Nyquist rate. The blue solid line represents the Nyquist-rate (analog) capacity, the black dotted line represents the capacity achieved by minimax sampler, the orange dashed illustrates the Nyquist-rate capacity minus the minimax capacity loss, while the purple dashed line corresponds to maximum worst-case capacity.

a desired minimax sampler controls the capacity loss for every single state $\boldsymbol{s}$, and allows for robustness over all channel states with universal channel-independent sampling. It turns out that in the compound multiband channels,

$$\forall \boldsymbol{s}, \quad L_{\boldsymbol{s}}^{\boldsymbol{Q}^{\mathrm{m}}} = \max_{\tilde{\boldsymbol{s}} \in \binom{[n]}{k}} L_{\tilde{\boldsymbol{s}}}^{\boldsymbol{Q}^{\mathrm{m}}}$$

except for some vanishingly small residual terms, as will be shown in the next section.

III. Minimax Sampled Capacity Loss

The minimax sampled capacity loss problem can be cast as minimizing $\max_{\boldsymbol{s}\in\mathcal{S}} L_{\boldsymbol{s}}^{\boldsymbol{Q}}$ over all sampling coefficient functions $\boldsymbol{Q}(f)$. In general, the problem is non-convex in $\boldsymbol{Q}(f)$, and hence it is difficult to identify the optimal sampling systems. Nevertheless, the minimax capacity loss can be quantified reasonably well at moderate-to-high SNR. It turns out that under both Landau-rate sampling and a large class of super-Landau-rate sampling, the minimax capacity loss can be approached arbitrarily well by random sampling.

Define the undersampling factor $\alpha := m/n$, and recall that the band sparsity ratio is $\beta := k/n$. Our main results are summarized in the following theorem.

Theorem 2. *Consider any sampling coefficient function $\boldsymbol{Q}(\cdot)$ with an undersampling factor α, and let the sparsity ratio be β. Define* $\mathrm{SNR}_{\min} := \frac{P}{\beta W} \inf_{0\leq f\leq W, \boldsymbol{s}\in\binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}$ *and* $\mathrm{SNR}_{\max} := \frac{P}{\beta W} \sup_{0\leq f\leq W, \boldsymbol{s}\in\binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}$.

(i) (Landau-rate sampling) If $\alpha = \beta$ (or $k = m$), then

$$\inf_{\boldsymbol{Q}} \max_{\boldsymbol{s}\in\binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{Q}} = \frac{W}{2}\left\{\mathcal{H}(\beta) + O\left(\frac{\log n}{n}\right) + \Delta_{\mathrm{L}}\right\}, \tag{7}$$

where

$$-\frac{2}{\sqrt{\mathrm{SNR}_{\min}}} \leq \Delta_{\mathrm{L}} \leq \frac{\beta}{\mathrm{SNR}_{\min}}.$$

(ii) (Super-Landau-rate sampling) Suppose that there is a small constant $\delta > 0$ such that $\alpha - \beta \geq \delta$ and $1 - \alpha - \beta \geq \delta$. Then

$$\inf_{\boldsymbol{Q}} \max_{\boldsymbol{s} \in \binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{Q}} = \frac{W}{2} \left\{ \mathcal{H}(\beta) - \alpha \mathcal{H}\left(\frac{\beta}{\alpha}\right) + O\left(\frac{\log^2 n}{\sqrt{n}}\right) + \Delta_{\mathrm{SL}} \right\}, \tag{8}$$

where

$$-\frac{2}{\sqrt{\mathrm{SNR}_{\min}}} \leq \Delta_{\mathrm{SL}} \leq \frac{\beta}{\mathrm{SNR}_{\min}}.$$

Remark 1. *Note that $\mathcal{H}(\cdot)$ denotes the entropy function. Its appearance is due to the fact that it is an asymptotically tight estimate of the rate function of binomial coefficients.*

Theorem 2 provides a tight characterization of the minimax sampled capacity loss relative to the Nyquist-rate capacity, under both Landau-rate sampling and super-Landau-rate sampling. Note that the Landau-rate sampling regime in (i) is not a special case of the super-Landau-rate regime considered in (ii). For instance, if $\beta > 1/2$, then $\alpha + \beta > 1$, which falls within a regime not accounted for by Theorem 2(ii).

The expressions (7) and (8) contain residual terms no larger than $O\left(\frac{\log n}{n}\right) + \frac{1}{\sqrt{\mathrm{SNR}_{\min}}}$ per unit bandwidth, which vanishes for large n and high SNR. These fundamental minimax limits do not scale with the SNR and n modulo a vanishing residual term. Since the Nyquist rate capacity scales as $\Theta(W \log \mathrm{SNR})$, our results indicate that the ratio of the minimax capacity loss to the Nyquist-rate capacity vanishes at a rate $\Theta(1/\log \mathrm{SNR})$.

Note that even if we allow power control at the transmitter side, the results are still valid at high SNR. This is summarized in the following theorem.

Theorem 3. *Consider the metric $L_{\boldsymbol{s}}^{\boldsymbol{Q},\mathrm{opt}}$ with power control. Under all other conditions of Theorem 2, the bounds (7) and (8) (with $L_{\boldsymbol{s}}^{\boldsymbol{Q}}$ replaced by $L_{\boldsymbol{s}}^{\boldsymbol{Q},\mathrm{opt}}$) continue to hold if Δ_{L} and Δ_{SL} are respectively replaced by some residuals $\Delta_{\mathrm{L}}^{\mathrm{opt}}$ and $\Delta_{\mathrm{SL}}^{\mathrm{opt}}$ that satisfy*

$$-\frac{2}{\sqrt{\mathrm{SNR}_{\min}}} \leq \Delta_{\mathrm{L}}^{\mathrm{opt}}, \Delta_{\mathrm{SL}}^{\mathrm{opt}} \leq \frac{\beta + \overline{A}}{\mathrm{SNR}_{\min}},$$

where $\overline{A}$ is a constant defined as

$$\overline{A} := \min \left\{ \frac{\max_{\boldsymbol{s} \in \binom{[n]}{k}} \int_0^W \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)} \mathrm{d}f}{\beta W \inf_{0 \leq f \leq W, \boldsymbol{s} \in \binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}}, \frac{\sup_{0 \leq f \leq W, \boldsymbol{s} \in \binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}}{\inf_{0 \leq f \leq W, \boldsymbol{s} \in \binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}} \right\}.$$

Theorem 3 demonstrates that if the average-to-minimum ratio $\frac{\overline{\mathrm{SNR}}}{\mathrm{SNR}_{\min}}$ is bounded by a constant (where $\overline{\mathrm{SNR}} := \max_{\boldsymbol{s} \in \binom{[n]}{k}} \frac{P}{\beta W} \int_0^W \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)} \mathrm{d}f$), then the minimax sampled capacity gap with power control remains almost the same as that with power control within a gap at most $O\left(\frac{\log n}{n}\right) + O\left(\frac{1}{\sqrt{\mathrm{SNR}_{\min}}}\right)$ per unit bandwidth. Note that the constant $\overline{A}$ given in Theorem 3 is fairly conservative, and can be refined with finer tuning or algebraic techniques.

Theorem 3 can be delivered as an immediate consequence of Theorem 2 if we can quantify the gap between $C_{\boldsymbol{s}}^{P_{\mathrm{eq}}}$ and $C_{\boldsymbol{s}}^{\mathrm{opt}}$. In fact, the capacity benefits of using power control at high SNR regime is no larger than $O\left(\frac{1}{\mathrm{SNR}}\right)$ per unit bandwidth. See Appendix A for details. For this reason, our analysis is mainly devoted to $L_{\boldsymbol{s}}^{\boldsymbol{Q}}$, which corresponds to the capacity loss relative to Nyquist-rate capacity with uniform power allocation.

The proof of Theorem 2 involves the verification of two parts: a converse part that provides a fundamental lower bound on the minimax sampled capacity loss, and an achievability part that provides a sampling scheme to approach this bound. As we show, the class of sampling systems with periodic modulation followed by low-pass filters, as illustrated in Fig. 1(b), is sufficient to approach the minimax sampled capacity loss.

Throughout the remainder of the paper, we suppose that the noise is of unit power spectral density $\mathcal{S}_\eta(f) \equiv 1$ unless otherwise specified. Note that this incurs no loss of generality since we can always include a noise-whitening LTI filter at the first stage of the sampling system.

A. The Converse

We need to show that the minimax sampled capacity loss under any channel-independent sampler cannot be lower than (7) and (8). This is given by the following theorem, which takes into account the entire regime including the situation where $\alpha + \beta > 1$.

Theorem 4. *Consider any Riemann-integrable sampling coefficient function $\boldsymbol{Q}(\cdot)$ with an undersampling factor $\alpha := m/n$. Suppose the sparsity ratio $\beta := k/n$ satisfies $\beta \leq \alpha \leq 1$. The minimax capacity loss can be lower bounded by*

$$\inf_{\boldsymbol{Q}} \max_{\boldsymbol{s}\in\binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{Q}} \geq \frac{W}{2}\left\{\mathcal{H}\left(\beta\right) - \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) - \frac{2}{\sqrt{\mathrm{SNR}_{\min}}} - \frac{\log\left(n+1\right)}{n}\right\}. \tag{9}$$

For a given β, the bound is decreasing in α. While the active channel bandwidth is smaller than the total bandwidth, the noise (even though the SNR is large) is scattered over the entire bandwidth. Thus, none of the universal sub-Nyquist sampling strategies are information preserving, and increasing the sampling rate can always harvest capacity gain.

B. Achievability with Landau-rate Sampling ($\alpha = \beta$)

Consider the achievability part when the sampling rate equals the active frequency bandwidth ($\beta = \alpha$). In general, it is very difficult to find a deterministic solution to approach the lower bound (9). A special instance of sampling methods that we can analyze concerns the case in which the sampling coefficient functions are flat over $[0, W/n]$ and whose coefficients are generated in a random fashion. It turns out that as n grows large, the capacity loss achievable by random sampling approaches the lower bound (9) uniformly across all realizations. The results are stated in Theorem 5 after introducing a class of sub-Gaussian measure below.

Definition 4. A measure ν on $\mathbb{R}$ satisfies the logarithmic Sobolev inequality with constant c_{LS} if, for any differentiable function g,

$$\int g^2 \log \frac{g^2}{\int g^2 \mathrm{d}\nu} \mathrm{d}\nu \leq 2c_{\mathrm{LS}} \int \left|g'\right|^2 \mathrm{d}\nu.$$

Remark 2. *A probability measure obeying the logarithmic Sobolev inequality possesses sub-Gaussian tails, and a large class of sub-Gaussian measures satisfies this inequality for some constant. See [29] for examples. In particular, the standard Gaussian measure satisfies this inequality with constant $c_{\mathrm{LS}} = 1$ (e.g. [33]).*

Theorem 5. *Let* $\boldsymbol{M} = (\zeta_{ij})_{1\leq i\leq k, 1\leq j\leq n}$ *be a random matrix in which* ζ_{ij}*'s are jointly independent with zero mean and unit variance. In addition, suppose that* ζ_{ij} *satisfies one of the following conditions:*

(a) ζ_{ij} *is almost surely bounded by a constant* D*;*

(b) The probability measure of ζ_{ij} *satisfies the logarithmic Sobolev inequality with a bounded constant* c_{LS}*.*

Then there exist some constants $c, C > 0$ *such that*

$$\max_{\boldsymbol{s}\in\binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{M}} \leq \frac{W}{2}\left(\mathcal{H}\left(\beta\right) + \frac{5\log k}{n} + \frac{\beta}{\mathrm{SNR}_{\min}}\right) \tag{10}$$

with probability exceeding $1 - C\exp\left(cn\right)$*.*

Theorem 5 demonstrates that independent random sampling achieves minimax sampled capacity loss among all periodic sampling methods with period n/W. In fact, our analysis demonstrates that the sampled capacity loss approaches the minimax limit uniformly over all states $\boldsymbol{s}$. Another interesting observation is the universality phenomenon, i.e. a broad class of sub-Gaussian ensembles, as long as the entries are jointly independent with matching moments, suffices to generate minimax samplers.

C. Achievability with Super-Landau-Rate Sampling $(\alpha > \beta, \alpha + \beta < 1)$

So far we have considered the case in which the sampling rate is equal to the spectral support. While the active bandwidth for transmission is smaller than the total bandwidth, the noise (even though the SNR is large) is scattered over the entire bandwidth. This indicates that none of the sub-Nyquist sampling strategies preserves all information contents conveyed through the noisy channel, unless they know the channel support. One may thus hope that increasing the sampling rate would improve the achievable information rate. The achievability result for super-Landau-rate sampling is stated in the following theorem.

Theorem 6. *Let* $\boldsymbol{M} = (\zeta_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ *be a random Gaussian matrix in which* ζ_{ij}*'s are i.i.d. drawn from* $\mathcal{N}(0,1)$*. Suppose that there exists a small constant* $\varepsilon > 0$ *such that* $1-\alpha-\beta \geq \varepsilon$ *and* $\alpha - \beta \geq \varepsilon$*. Then there exist some constants* $c, C > 0$ *such that*

$$\max_{\boldsymbol{s}\in\binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{M}} \leq \frac{W}{2}\left[\mathcal{H}(\beta) - \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) + O\left(\frac{\log^2 n}{\sqrt{n}}\right) + \frac{\beta}{\mathrm{SNR}_{\min}}\right]$$

with probability exceeding $1 - C\exp\left(-cn\right)$*.*

Theorem 6 indicates that i.i.d. Gaussian sampling approaches the minimax capacity loss (8) with vanishingly small gap. As will be shown in our proof, with exponentially high probability, the sampled capacity loss for all states are equivalent and coincide with the fundamental minimax limit. In contrast to Theorem 5, we restrict our attention to Gaussian sampling, which suffices for the proof of Theorem 2.

D. Equivalent Algebraic Problems

Our results are established by investigating three equivalent algebraic problems. Recall that $\frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}^2 \succeq \mathrm{SNR}_{\min}\boldsymbol{I}_k$. Define $\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}} := (\boldsymbol{Q}\boldsymbol{Q}^*)^{-\frac{1}{2}}\boldsymbol{Q}_{\boldsymbol{s}}$. Simple manipulations yield

$$\begin{aligned}
L_{\boldsymbol{s}}^{\boldsymbol{Q}} &= -\frac{1}{2}\int_0^{W/n} \log\det\left(\boldsymbol{I}_m + \frac{P}{\beta W}\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}(f)\boldsymbol{H}_{\boldsymbol{s}}^2(f)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}(f)\right)\mathrm{d}f + \frac{1}{2}\int_0^{W/n}\log\det\left(\boldsymbol{I}_k + \frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}^2(f)\right)\mathrm{d}f \\
&= -\frac{1}{2}\int_0^{W/n} \log\det\left(\boldsymbol{I}_k + \frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}(f)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}(f)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}(f)\boldsymbol{H}_{\boldsymbol{s}}(f)\right)\mathrm{d}f \\
&\qquad + \frac{1}{2}\int_0^{W/n}\log\det\left(\frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}^2(f)\right)\mathrm{d}f + \frac{\beta W}{2}\Delta_{\boldsymbol{s}} \qquad (11) \\
&= -\frac{1}{2}\int_0^{W/n} \log\det\left(\frac{\beta W}{P}\boldsymbol{H}_{\boldsymbol{s}}^{-2}(f) + \boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}(f)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}(f)\right)\mathrm{d}f + \frac{\beta W}{2}\Delta_{\boldsymbol{s}}, \qquad (12)
\end{aligned}$$

where $\Delta_{\boldsymbol{s}}$ denotes some residual term. In particular, $\Delta_{\boldsymbol{s}}$ can be bounded as

$$0 \le \Delta_{\boldsymbol{s}} \le \frac{1}{\mathrm{SNR}_{\min}}. \qquad (13)$$

This is an immediate consequence of the following observation: for any $k \times k$ positive semidefinite matrix $\boldsymbol{A}$,

$$0 \le \frac{1}{k}\log\det(\boldsymbol{I}_k + \boldsymbol{A}) - \frac{1}{k}\log\det(\boldsymbol{A}) = \frac{1}{k}\sum_{i=1}^{k}\log\left(1 + \frac{1}{\lambda_i(\boldsymbol{A})}\right) \le \frac{1}{\lambda_{\min}(\boldsymbol{A})}. \qquad (14)$$

Recall that $\mathrm{SNR}_{\min} := \frac{P}{\beta W}\inf_{0\le f\le W}|H(f)|^2$ and $\mathrm{SNR}_{\max} := \frac{P}{\beta W}\sup_{0\le f\le W}|H(f)|^2$. Therefore, $\frac{\beta W}{P}\boldsymbol{H}_{\boldsymbol{s}}^{-2}$ can be bounded as

$$\frac{1}{\mathrm{SNR}_{\min}}\boldsymbol{I}_k \preceq \frac{\beta W}{P}\boldsymbol{H}_{\boldsymbol{s}}^{-2} \preceq \frac{1}{\mathrm{SNR}_{\min}}\boldsymbol{I}_k. \qquad (15)$$

This bound together with (12) makes $\det\left(\epsilon\boldsymbol{I}_k + \boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}\right)$ a quantity of interest (for some small ϵ).

1) The Converse: Note that $\boldsymbol{Q}^{\mathrm{w}}(f)$ has orthonormal rows. The following theorem investigates the properties of $\det(\epsilon\boldsymbol{I}_k + \boldsymbol{B}_{\boldsymbol{s}}^*\boldsymbol{B}_{\boldsymbol{s}})$ for any $m\times n$ matrix $\boldsymbol{B}$ that has orthonormal rows. This, together with the Riemann integrability assumption of $\boldsymbol{Q}(f)$, establishes Theorem 4.

Theorem 7. *(1) Consider any $m \times n$ matrix $\boldsymbol{B}$ $(n \ge m \ge k)$ that satisfies $\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{I}_m$, and denote by $\boldsymbol{B}_{\boldsymbol{s}}$ the $m \times k$ submatrix of $\boldsymbol{B}$ with columns coming from the index set $\boldsymbol{s}$. Then for any $\epsilon > 0$, one has*

$$\begin{aligned}
\binom{m}{k} \le \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \det(\epsilon\boldsymbol{I}_m + \boldsymbol{B}_{\boldsymbol{s}}^*\boldsymbol{B}_{\boldsymbol{s}}) &= \sum_{l=0}^{k}\binom{n-l}{k-l}\binom{m}{l}\epsilon^{k-l} \qquad (16) \\
&\le \binom{m}{k}\left(1+\sqrt{\epsilon}\right)^{n+k}. \qquad (17)
\end{aligned}$$

(2) For any positive integer p, suppose that $\boldsymbol{B}_1, \cdots, \boldsymbol{B}_p$ are all $m \times n$ matrices such that $\boldsymbol{B}_i\boldsymbol{B}_i^* = \boldsymbol{I}_m$. Then,

$$\begin{aligned}
\min_{\boldsymbol{s}\in\binom{[n]}{k}} \frac{1}{np}\sum_{i=1}^{p}\log\det\left(\epsilon\boldsymbol{I}_k + (\boldsymbol{B}_i)_{\boldsymbol{s}}^*(\boldsymbol{B}_i)_{\boldsymbol{s}}\right) &\le \frac{1}{n}\log\binom{m}{k} - \frac{1}{n}\log\binom{n}{k} + 2\sqrt{\epsilon} \qquad (18) \\
&\le \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) - \mathcal{H}(\beta) + 2\sqrt{\epsilon} + \frac{\log(n+1)}{n}. \qquad (19)
\end{aligned}$$

Note that $\boldsymbol{Q}^{\mathrm{w}}(f)$ has orthonormal rows for any f, and $\boldsymbol{Q}^{\mathrm{w}}(f)$ is assumed to be Riemann integrable. For any $\delta > 0$, we can find a sufficiently large p such that

$$\int_0^{W/n} \log\det\left(\epsilon \boldsymbol{I}_k + \boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}(f)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}(f)\right)\mathrm{d}f \leq \delta + \frac{W}{pn}\sum_{i=1}^{p}\log\det\left(\epsilon \boldsymbol{I}_k + \boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}\left(\frac{iW}{pn}\right)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}\left(\frac{iW}{pn}\right)\right).$$

Since δ can be arbitrarily small, applying Theorem 7 immediately yields that for any $\boldsymbol{Q}(\cdot)$:

$$\min_{\boldsymbol{s}\in\binom{[n]}{k}} \int_0^{W/n} \log\det\left(\epsilon \boldsymbol{I}_k + \boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}(f)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}(f)\right)\mathrm{d}f \leq W\left\{\alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) - \mathcal{H}\left(\beta\right) + 2\sqrt{\epsilon} + \frac{\log\left(n+1\right)}{n}\right\}.$$

This together with (12), (13) and (15) leads to

$$\inf_{\boldsymbol{Q}} \max_{\boldsymbol{s}\in\binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{Q}} \geq \frac{W}{2}\left\{\mathcal{H}\left(\beta\right) - \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) - \frac{2}{\sqrt{\mathrm{SNR}_{\min}}} - \frac{\log\left(n+1\right)}{n}\right\},$$

which completes the proof of Theorem 4.

2) Achievability (Landau-rate Sampling): When it comes to the achievability part, the major step is to quantify $\det(\epsilon\boldsymbol{I}_k + (\boldsymbol{M}\boldsymbol{M}^T)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T)$ for every $\boldsymbol{s}$. Interestingly, this quantity can be uniformly bounded due to the concentration of spectral measure of random matrices [29]. This is stated in the following theorem, which demonstrates the achievability for Landau-rate sampling.

Theorem 8. *Let $\boldsymbol{M} = (\zeta_{ij})_{1\leq i\leq k, 1\leq j\leq n}$ be a real-valued random matrix. Under the conditions of Theorem 4, one has*

$$-\mathcal{H}\left(\beta\right) - \frac{5\log k}{n} \leq \min_{\boldsymbol{s}\in\binom{[n]}{k}} \frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k + \left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right) \leq -\mathcal{H}\left(\beta\right) + 2\sqrt{\epsilon} + \frac{\log\left(n+1\right)}{n}$$

with probability exceeding $1 - C\exp\left(-cn\right)$ for some constants $c, C > 0$.

Putting Theorem 8 and equations (12), (13) and (15) together establishes that

$$\max_{\boldsymbol{s}\in\binom{[n]}{k}} L_{\boldsymbol{s}}^{\boldsymbol{M}} \leq \frac{W}{2}\left[\mathcal{H}\left(\beta\right) + \frac{5\log k}{n} + \frac{\beta}{\mathrm{SNR}_{\min}}\right]$$

with probability exceeding $1 - C\exp\left(cn\right)$.

3) Achievability (Super-Landau-rate Sampling): Instead of studying a large class of sub-Gaussian random ensembles[2], the following theorem focuses on i.i.d. Gaussian matrices, which establishes the optimality of Gaussian random sampling for the super-Landau regime.

Theorem 9. *Let $\boldsymbol{M} = (\zeta_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ be a real-valued i.i.d. Gaussian matrix. Under the conditions of Theorem 6, one has*

$$\begin{aligned}-\mathcal{H}(\beta) + \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) + O\left(\frac{\log^2 n}{\sqrt{n}}\right) &\leq \min_{\boldsymbol{s}\in\binom{[n]}{k}} \frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k + \boldsymbol{M}_{\boldsymbol{s}}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right) \\ &\leq -\mathcal{H}\left(\beta\right) + \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) + 2\sqrt{\epsilon} + \frac{\log\left(n+1\right)}{n}\end{aligned}$$

[2]The proof argument for Landau-rate sampling cannot be readily carried over to super-Landau regime since $\boldsymbol{M}_{\boldsymbol{s}}$ is now a tall matrix, and hence we cannot separate $\boldsymbol{M}_{\boldsymbol{s}}$ and $\boldsymbol{M}\boldsymbol{M}^*$ easily.

with probability at least $1-C\exp\left(-cn\right)$ *for some constants* $c,C>0$.

Combining Theorem 9 and equations (12), (13) and (15) implies that

$$L_{\boldsymbol{s}}^{\boldsymbol{M}}\leq\frac{W}{2}\left[\mathcal{H}(\beta)-\alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right)+O\left(\frac{\log^{2}n}{\sqrt{n}}\right)+\frac{\beta}{\mathrm{SNR}_{\min}}\right]$$

with probability exceeding $1-C\exp\left(-cn\right)$.

The proofs of Theorems 7, 8 and 9, which are provided in Appendices B, C and D, respectively, rely heavily on *non-asymptotic* (random) matrix theory.

IV. Discussion

A. Implications of Main Results

Under both Landau-rate and super-Landau-rate sampling, the minimax capacity loss depends almost entirely on β and α. In this subsection, we summarize several key insights from the main theorems.

1) The Converse: Our analysis demonstrates that at high SNR, the loss $L_{\boldsymbol{s}}^{\boldsymbol{Q}}$ depends almost solely on the quantity

$$d\left(\boldsymbol{Q}(f),\boldsymbol{s},\epsilon\right):=\det\left(\epsilon\boldsymbol{I}_{k}+\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}*}(f)\boldsymbol{Q}_{\boldsymbol{s}}^{\mathrm{w}}(f)\right)$$

for small $\epsilon>0$, which is approximately the exponential of capacity loss at a given pair $(\boldsymbol{s},f)$. In fact, the key observation underlying the proof of Theorem 7 is that for any f, the sum

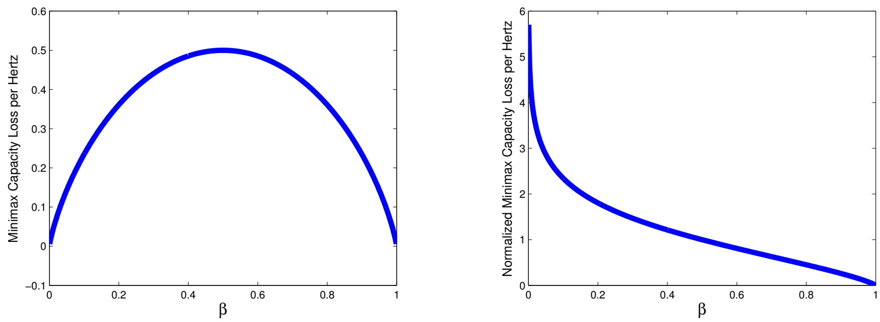
$$\sum_{\boldsymbol{s}\in\binom{[n]}{k}}d\left(\boldsymbol{Q}(f),\boldsymbol{s},\epsilon\right)$$

is a *constant* independent of the sampling coefficient function $\boldsymbol{Q}$. In other words, at any given f, the exponential sum of capacity loss over all states $\boldsymbol{s}$ is invariable regardless of what samplers we employ.

This invariant quantity is critical in identifying the minimax sampling method. In fact, it motivates us to seek a sampling method that achieves equivalent performance over all states $\boldsymbol{s}$. Large random matrices exhibit sharp concentration of spectral measure, and hence become a natural candidate to attain minimaxity.

2) Landau-rate Sampling: When sampling is performed at the Landau rate, the minimax capacity loss per unit Hertz is almost solely determined by the entropy function $\mathcal{H}(\beta)$. Specifically, when n and k are sufficiently large, the minimax limit depends only on the sparsity ratio $\beta=\frac{k}{n}$ rather than (n,k). Some implications from Theorem 4 and Theorem 5 are as follows.

1) The capacity loss per unit Hertz is illustrated in Fig. 4(a). The capacity loss vanishes when $\beta\rightarrow1$, since Nyquist-rate sampling is information preserving. The capacity loss divided by β is plotted in Fig. 4(b), which provides a normalized view of the capacity loss. It can be seen that the normalized loss decreases monotonically with β, indicating that the loss is more severe in sparse channels. Note that this is different from an LTI channel whereby sampling at the Landau rate is sufficient to preserve all information. When the channel state is uncertain, increasing the sampling rate above the Landau rate (but below the Nyquist rate) effectively increases the SNR, and hence allows more information to be harvested from the noisy sampled output.



(a) (b)

Figure 4. Plot (a) illustrates $\mathcal{H}(\beta)/2$ v.s. the sparsity ratio β, which characterizes the fundamental minimax capacity loss pert Hertz within a gap at most $O\left(\frac{\log n}{n}\right)+\frac{2}{\sqrt{\text{SNR}_{\min}}}$. Plot (b) illustrates $\mathcal{H}(\beta)/2\beta$ v.s. β, which corresponds approximately to the normalized capacity loss per Hertz.

2) The capacity loss incurred by independent random sampling meets the fundamental minimax limit for Landau-rate sampling uniformly across all states $\boldsymbol{s}$, which reveals that with exponentially high probability, random sampling is optimal in terms of universal sampling design. The capacity achievable by random sampling exhibits very sharp concentration around the minimax limit uniformly across all states $\boldsymbol{s}\in\binom{[n]}{k}$.
3)
4) A *universality* phenomenon that arises in large random matrices (e.g. [34]) leads to the fact that the minimaxity of random sampling matrices does not depend on the particular distribution of the coefficients. For a large class of sub-Gaussian measure, as long as all entries are jointly independent with matching moments up to the second order, the sampling mechanism it generates is minimax with exponentially high probability.

3) Super-Landau-Rate Sampling: The random sampling analyzed in Theorem 6 only involves Gaussian random sampling, and we have not shown universality results. Some implications under super-Landau-rate sampling are as follows.

1) Similar to the case with Landau-rate sampling, Gaussian sampling achieves minimax capacity loss uniformly across all states $\boldsymbol{s}$ within a large super-Landau-rate regime. The capacity gap is illustrated in Fig. 5. It can be observed from the plot that increasing the α/β ratio improves the capacity gap, shrinks the locus and shifts it leftwards.
2) Theorem 6 only concerns i.i.d. Gaussian random sampling instead of more general independent random sampling. While we conjecture that the universality phenomenon continues to hold for other jointly independent random ensembles with sub-Gaussian tails and appropriate matching moments, the mathematical analysis turns out to be more tricky than in the Landau-rate sampling case.
3) The capacity gain by sampling above the Landau rate depends on the undersampling factor α as well. Specifically, the capacity benefit per unit bandwidth due to super-Landau sampling is captured by the term

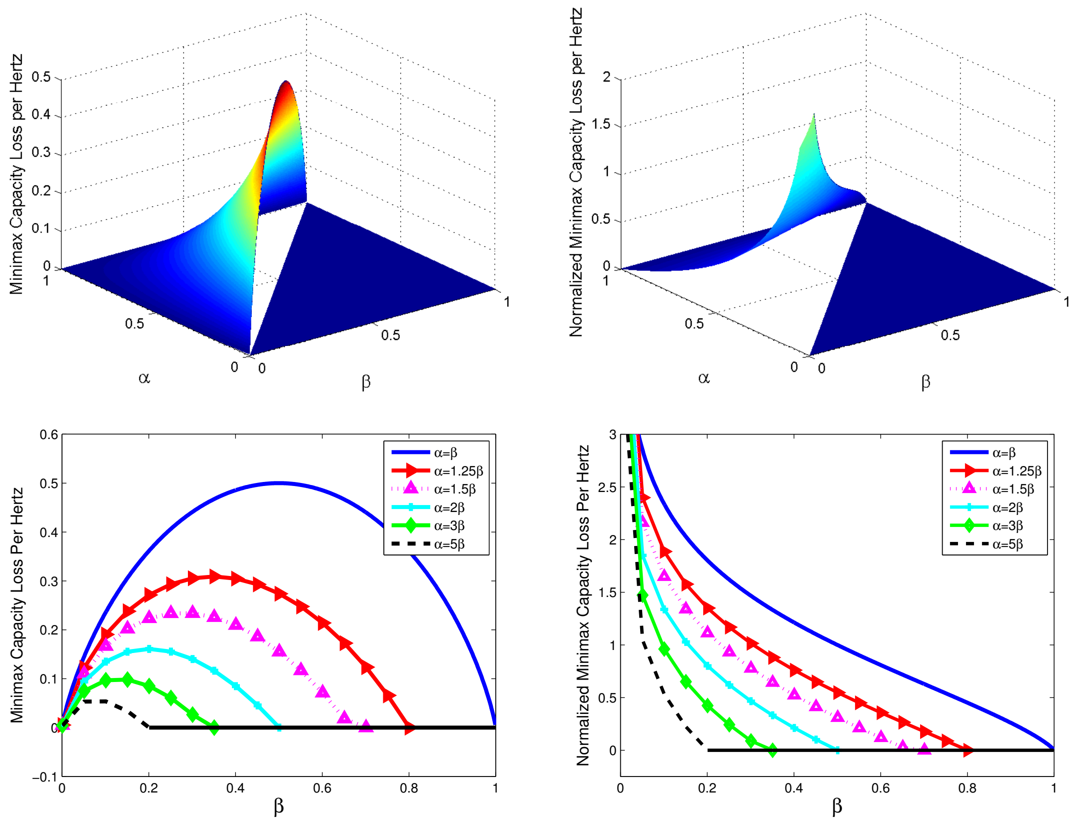


Figure 5. The function $\frac{1}{2}\mathcal{H}(\beta) - \frac{\alpha}{2}\mathcal{H}(\beta/\alpha)$ v.s. the sparsity ratio β and the undersampling factor α. Here, $\frac{1}{2}\mathcal{H}(\beta) - \frac{\alpha}{2}\mathcal{H}(\beta/\alpha)$ characterizes the fundamental minimax capacity loss pert Hertz within a gap at most $O\left(\frac{\log n}{n}\right) + \frac{2}{\sqrt{\mathrm{SNR}_{\min}}}$.

$\frac{1}{2}\alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right)$. When $\alpha \to 1$, the capacity loss per Hertz reduces to

$$\frac{1}{2}\mathcal{H}(\beta) - \frac{1}{2}\alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) = 0,$$

meaning that there is effectively no capacity loss under Nyquist-rate sampling. This agrees with the fact that Nyquist-rate sampling is information preserving.

V. Connections with Discrete-Time Sparse Channels

A. *Minimax Sampler in Discrete-time Sparse Channels*

Our results have straightforward extensions to discrete-time sparse vector channels. Specifically, consider a collection of n parallel channels. The channel input $\boldsymbol{x} \in \mathbb{R}^n$ is passed though the channel and contaminated by Gaussian noise $\boldsymbol{n} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, yielding a channel output

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n},$$

where $\boldsymbol{H}$ is some diagonal channel matrix. In particular, at each asymptotically long timeframe, a state $\boldsymbol{s} \in \binom{[n]}{k}$ is generated, which dictates the set of channels available for transmission, i.e. the transmitter can only send $\boldsymbol{x}$ at

indices in $\boldsymbol{s}$. One can then obtain m measurements of the channel output through a sensing matrix $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$, i.e. the measurements $\boldsymbol{y} \in \mathbb{R}^{m}$ can be expressed as

$$\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{r} = \boldsymbol{Q}\left(\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}\right).$$

The goal is then to identify a sensing matrix $\boldsymbol{Q}$ that minimizes the worst-case capacity loss over all states $\boldsymbol{s} \in \binom{[n]}{k}$.

If we abuse our notation $\boldsymbol{L}_{\boldsymbol{s}}^{\boldsymbol{Q}}$ (resp. $\boldsymbol{L}_{\boldsymbol{s}}^{\boldsymbol{Q},\mathrm{opt}}$) to denote the capacity loss at state $\boldsymbol{s}$ relative to Nyquist-rate capacity without (resp. with) power control, then the following results are immediate.

Theorem 10. *Define* $\mathrm{SNR}_{\min} := \frac{P}{k}\inf_{1\leq i\leq n}|\boldsymbol{H}_{ii}|$ *and* $\mathrm{SNR}_{\max} := \frac{P}{k}\sup_{1\leq i\leq n}|\boldsymbol{H}_{ii}|$.

(i) (Landau-rate sampling) If $\alpha=\beta$ $(k=m)$*, then*

$$\inf_{\boldsymbol{Q}}\max_{\boldsymbol{s}\in\binom{[n]}{k}}\frac{L_{\boldsymbol{s}}^{\boldsymbol{Q}}}{n} = \frac{1}{2}\left\{\mathcal{H}\left(\beta\right) + O\left(\frac{\log n}{n}\right) + \Delta_{\mathrm{L}}\right\}, \tag{20}$$

and

$$\inf_{\boldsymbol{Q}}\max_{\boldsymbol{s}\in\binom{[n]}{k}}\frac{L_{\boldsymbol{s}}^{\boldsymbol{Q},\mathrm{opt}}}{n} = \frac{1}{2}\left\{\mathcal{H}\left(\beta\right) + O\left(\frac{\log n}{n}\right) + \Delta_{\mathrm{L}}^{\mathrm{wf}}\right\}. \tag{21}$$

(ii) (Super-Landau-rate sampling) Suppose that there is a small constant $\delta>0$ *such that* $\alpha-\beta\geq\delta$ *and* $1-\alpha-\beta\geq\delta$*. Then*

$$\inf_{\boldsymbol{Q}}\max_{\boldsymbol{s}\in\binom{[n]}{k}}\frac{L_{\boldsymbol{s}}^{\boldsymbol{Q}}}{n} = \frac{1}{2}\left\{\mathcal{H}\left(\beta\right) - \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) + O\left(\frac{\log^2 n}{\sqrt{n}}\right) + \Delta_{\mathrm{SL}}\right\}, \tag{22}$$

and

$$\inf_{\boldsymbol{Q}}\max_{\boldsymbol{s}\in\binom{[n]}{k}}\frac{L_{\boldsymbol{s}}^{\boldsymbol{Q},\mathrm{opt}}}{n} = \frac{1}{2}\left\{\mathcal{H}\left(\beta\right) - \alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right) + O\left(\frac{\log^2 n}{\sqrt{n}}\right) + \Delta_{\mathrm{SL}}^{\mathrm{wf}}\right\}. \tag{23}$$

Here, Δ_{L}*,* $\Delta_{\mathrm{L}}^{\mathrm{opt}}$*,* Δ_{SL} *and* $\Delta_{\mathrm{SL}}^{\mathrm{opt}}$ *are some residual terms satisfying*

$$-\frac{2}{\sqrt{\mathrm{SNR}_{\min}}} \leq \Delta_{\mathrm{L}}, \Delta_{\mathrm{SL}} \leq \frac{\beta}{\mathrm{SNR}_{\min}}, \quad \textit{and} \quad -\frac{2}{\sqrt{\mathrm{SNR}_{\min}}} \leq \Delta_{\mathrm{L}}^{\mathrm{opt}}, \Delta_{\mathrm{SL}}^{\mathrm{opt}} \leq \frac{\beta+\overline{A}}{\mathrm{SNR}_{\min}},$$

where $\overline{A}$ *is a constant defined as*

$$\overline{A} := \min\left\{\frac{\mathrm{tr}\left(\boldsymbol{H}\boldsymbol{H}^{*}\right)}{k\min_{1\leq i\leq n}\left|\boldsymbol{H}_{ii}^{2}\right|}, \frac{\max_{1\leq i\leq n}\left|\boldsymbol{H}_{ii}^{2}\right|}{\min_{1\leq i\leq n}\left|\boldsymbol{H}_{ii}^{2}\right|}\right\}.$$

The key observations are that in a discrete-time sparse vector channel, the minimax capacity loss per degree of freedom again depends only on β and α modulo a vanishingly small gap. Independent random sensing matrices and i.i.d. Gaussian sensing matrices are minimax in terms of a channel-blind sensing matrix design, in the Landau-rate and super-Landau-rate regimes, respectively.

B. Connections with Sparse Recovery

1) Restricted Isometry Property (RIP): The readers familiar with compressed sensing [7], [8] may naturally wonder whether the optimal sampling matrices $\boldsymbol{M}$ satisfy the restricted isometry property (RIP). An RIP constant δ_k with respect to a matrix $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ is defined (e.g. [35]) as the smallest quantity such that

$$\left(1-\delta_k\right)\|\boldsymbol{c}\| \leq \|\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{c}\|^2 \leq \left(1+\delta_k\right)\|\boldsymbol{c}\|$$

holds for any vector $\boldsymbol{c}$ and any $\boldsymbol{s}$ of size at most k (recall that $\boldsymbol{M}_{\boldsymbol{s}}$ is a submatrix consisting of k columns of $\boldsymbol{M}$). This quantity measures how close $\boldsymbol{M}_{\boldsymbol{s}}$'s are to orthonormal systems, and the existence of a small RIP constant that does not scale with (n, k, m) typically enables exact sparse recovery from noiseless measurements with a sensing matrix $\boldsymbol{M}$ [35].

Nevertheless, RIP is not necessary for approaching the minimax capacity loss. Consider the Landau-rate sampling regime for example. When the entries of $\boldsymbol{M}$ are independently generated under conditions of Theorem 8, one typically has (e.g. [36])

$$\sigma_{\min}\left(\boldsymbol{M}_{\boldsymbol{s}}\right) = O\left(\frac{1}{k}\right),$$

which cannot be bounded away from 0 by a constant. On the other hand, there is no guarantee that a restricted isometric matrix $\boldsymbol{M}$ is sufficient to achieve minimaxity. An optimal sampling matrix typically has similar spectrum as an i.i.d. Gaussian matrix, but a general restricted isometric matrix does not necessarily exhibit similar spectrum.

We note, however, that many randomized schemes for generating a restricted isometric matrix are natural candidates for generating minimax samplers. As shown by our analysis, in order to obtain a desired sampling matrix $\boldsymbol{M}$, we require $\boldsymbol{M}_{\boldsymbol{s}}$ to yield similar spectrum over all $\boldsymbol{s}$. Many random matrices satisfying RIP exhibit this property, and are hence minimax.

2) Necessary Sampling Rate: It is well known that there exists a spectrum-blind sampling matrix with $2k$ *noiseless* measurements that admits perfect recovery of any k-sparse signal. In the continuous-time counterpart, the minimum sampling rate for perfect recovery in the absence of noise is twice the spectrum occupancy [37]. Nevertheless, this sampling rate does not allow zero capacity loss in our setting. Since the channel output is contaminated by noise and thus has bandwidth W, a spectrum-blind sampler is unable to suppress spectral contents outside active subbands, and hence suffers from information rate loss relative to Nyquist-rate capacity.

On the other hand, twice the spectrum occupancy does not have a threshold effect in our setting, as illustrated in Figure 5. This arises from the fact that the transmitter can adapt its transmitted signal to the instantaneous realization and sampling rate. For instance, the spectral support of the transmitted signal may also shrink as the sampling rate decreases, thus avoiding an infliction point on the capacity curves.

VI. Conclusions

We have investigated optimal universal sampling design from a capacity perspective. In order to evaluate the loss due to universal sub-Nyquist sampling design, we introduced the notion of sampled capacity loss relative to Nyquist-rate capacity, and characterize overall robustness of the sampling design through the minimax capacity loss metric. Specifically, we have determined the fundamental minimax limit on the sampled capacity loss achievable by a class of channel-blind periodic sampling system. This minimax limit turns out to be a constant that only depends on the band sparsity ratio and undersampling factor, modulo a residual term that vanishes in the SNR and the number of subbands. Our results demonstrate that with exponentially high probability, random sampling is minimax in terms of a universal sampler design. This highlights the power of random sampling methods in the

channel-blind design. In addition, our results extend to discrete-time counterparts without difficulty. We demonstrate that independent random sensing matrices are minimax in discrete-time sparse vector channels.

It remains to characterize the fundamental minimax capacity loss when sampling is performed below the Landau rate, and to be seen whether random sampling is still optimal in the sub-Landau-rate regime. It would also be interesting to extend this framework to situations beyond compound multiband channels, and our notion of sampled capacity loss will be useful in evaluating the robustness for these scenarios. Our framework and results may also be appropriate for other channels with state where sparsity exists in other transform domains. In addition, when it comes to multiple access channels or random access channels [23], it is not clear how to find a channel-blind sampler that is robust for the entire capacity region.

APPENDIX A
PROOF OF THEOREM 3

We would like to bound the gap between $C_{\boldsymbol{s}}^{\mathrm{wf}}$ and $C_{\boldsymbol{s}}^{\mathrm{eq}}$. In fact, the equation that determines the water level implies that

$$P=\int_{0}^{W/n}\sum_{i=1}^{k}\left(\nu-\frac{1}{\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{ii}^{2}}\right)^{+}\mathrm{d}f\geq\int_{0}^{W/n}\sum_{i=1}^{k}\left(\nu-\frac{1}{\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{ii}^{2}}\right)\mathrm{d}f \tag{24}$$

which in turns yields

$$\nu\leq\frac{P}{\beta W}+\frac{\int_{0}^{W/n}\sum_{i=1}^{k}\frac{1}{\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{ii}^{2}}\mathrm{d}f}{\beta W}.$$

With the above bound on the water level, the capacity can be bounded above as

$$\begin{aligned}
C_{\boldsymbol{s}}^{\mathrm{opt}} &= \int_{0}^{W/n}\frac{1}{2}\sum_{i=1}^{k}\log^{+}\left(\nu\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{ii}^{2}\right)\mathrm{d}f\\
&\leq\int_{0}^{W/n}\frac{1}{2}\sum_{i=1}^{k}\log^{+}\left(\frac{\int_{0}^{W/n}\sum_{j=1}^{k}\frac{\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{ii}^{2}}{\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{jj}^{2}}\mathrm{d}f}{\beta W}+\frac{P}{\beta W}\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{ii}^{2}\right)\mathrm{d}f\\
&\leq\int_{0}^{W/n}\frac{1}{2}\sum_{i=1}^{k}\log\left(A+\frac{P}{\beta W}\left(\boldsymbol{H}_{\boldsymbol{s}}(f)\right)_{ii}^{2}\right)\mathrm{d}f\\
&=\int_{0}^{W/n}\log\det\left(A\boldsymbol{I}+\frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}^{2}(f)\right)\mathrm{d}f,
\end{aligned}$$

where $A := \max_{\boldsymbol{s},i} \frac{\int_0^{W/n} \sum_{j=1}^{k} \frac{(\boldsymbol{H_s}(f))_{ii}^2}{(\boldsymbol{H_s}(f))_{jj}^2} \mathrm{d}f}{\beta W}$. One can easily verify that $A \geq 1$. Therefore,

$$
\begin{aligned}
C_{\boldsymbol{s}}^{\text{opt}} - C_{\boldsymbol{s}}^{P_{\text{eq}}} &\leq \int_0^{W/n} \log\det\left(A\boldsymbol{I} + \frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}^2(f)\right)\mathrm{d}f - \int_0^{W/n} \log\det\left(\boldsymbol{I} + \frac{P}{\beta W}\boldsymbol{H}_{\boldsymbol{s}}^2(f)\right)\mathrm{d}f \\
&\leq \int_0^{W/n} \sum_{i=1}^{k} \log\frac{A + \frac{P}{\beta W}\left(\boldsymbol{H}_{\boldsymbol{s}}^2(f)\right)_{ii}}{1 + \frac{P}{\beta W}\left(\boldsymbol{H}_{\boldsymbol{s}}^2(f)\right)_{ii}}\mathrm{d}f \leq \int_0^{W/n} \sum_{i=1}^{k} \log\frac{A + \inf_{0\leq f\leq W} \frac{P}{\beta W}\frac{|H(f,\boldsymbol{s})|}{\sqrt{\mathcal{S}_\eta(f)}}}{1 + \inf_{0\leq f\leq W} \frac{P}{\beta W}\frac{|H(f,\boldsymbol{s})|}{\sqrt{\mathcal{S}_\eta(f)}}}\mathrm{d}f \\
&\leq \beta W \log\left(1 + \frac{A-1}{1+\mathsf{SNR}_{\min}}\right) \\
&\leq W\frac{\beta\,(A-1)}{1+\mathsf{SNR}_{\min}}.
\end{aligned}
$$

We also observe that

$$
\begin{aligned}
A = \max_{\boldsymbol{s},i} \frac{\int_0^{W/n} \sum_{j=1}^{k} \frac{(\boldsymbol{H_s}(f))_{ii}^2}{(\boldsymbol{H_s}(f))_{jj}^2}\mathrm{d}f}{\beta W} &\leq \frac{\int_0^{W/n} \sum_{j=1}^{k} \frac{(\boldsymbol{H_s}(f))_{ii}^2}{\inf_{f,\boldsymbol{s}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}}\mathrm{d}f}{\beta W} \\
&\leq \min\left\{ \frac{\max_{\boldsymbol{s}\in\binom{[n]}{k}} \int_0^W \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}\mathrm{d}f}{\beta W \inf_{0\leq f\leq W,\boldsymbol{s}\in\binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}}, \frac{\sup_{0\leq f\leq W,\boldsymbol{s}\in\binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}}{\inf_{0\leq f\leq W,\boldsymbol{s}\in\binom{[n]}{k}} \frac{|H(f,\boldsymbol{s})|^2}{\mathcal{S}_\eta(f)}} \right\}.
\end{aligned}
$$

Combining the above bounds and Theorem 2 completes the proof.

APPENDIX B
PROOF OF THEOREM 7

Before proving the results, we first state two facts. Consider any $m \times m$ matrix $\boldsymbol{A}$, and list the eigenvalues of $\boldsymbol{A}$ as $\lambda_1, \cdots, \lambda_m$. Define the characteristic polynomial of $\boldsymbol{A}$ as

$$p_{\boldsymbol{A}}(t) = \det(t\boldsymbol{I} - \boldsymbol{A}) = t^m - S_1(\lambda_1, \cdots, \lambda_m)t^{m-1} + \cdots + (-1)^m S_m(\lambda_1, \cdots, \lambda_m),$$

where $S_l(\lambda_1, \cdots, \lambda_m)$ is the lth elementary symmetric function of $\lambda_1, \cdots, \lambda_m$ defined as follows:

$$S_l(\lambda_1, \cdots, \lambda_m) := \sum_{1\leq i_1 < \cdots < i_l \leq m} \prod_{j=1}^{l} \lambda_{i_j}.$$

We also define $E_l(\boldsymbol{A})$ as the sum of determinants of all l-by-l principal minors of $\boldsymbol{A}$. According to [38, Theorem 1.2.12], $S_l(\lambda_1, \cdots, \lambda_m) = E_l(\boldsymbol{A})$ holds for all $1 \leq l \leq m$. After a little manipulation we obtain

$$\det(t\boldsymbol{I} + \boldsymbol{A}) = t^m + E_1(\boldsymbol{A})t^{m-1} + \cdots + E_k(\boldsymbol{A}). \tag{25}$$

Another fact we would like to stress is the entropy formula of binomial coefficients. Specifically, for any $0 < k < n$, one has [39, Page 43]

$$\frac{e^{n\mathcal{H}(\beta)}}{n+1} \leq \binom{n}{k} \leq e^{n\mathcal{H}(\beta)},$$

and hence

$$\mathcal{H}(\beta) - \frac{\log(n+1)}{n} \leq \frac{1}{n}\log\binom{n}{k} \leq \mathcal{H}(\beta), \tag{26}$$

where $\mathcal{H}(x) := x\log\frac{1}{x} + (1-x)\log\frac{1}{1-x}$ denotes the entropy function.

Now we are in a position to derive the proof for our main results.

(1) Consider an $m \times n$ matrix $\boldsymbol{B}$ with orthonormal rows ($m \geq k$), i.e. $\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{I}_m$. Using this identity (25), we can derive

$$\sum_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I}_m + \boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right) = \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \left\{\epsilon^m + \sum_{l=1}^{m} \epsilon^{m-l} E_l\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right)\right\}$$
$$= \epsilon^m \binom{n}{k} + \sum_{l=1}^{k} \epsilon^{m-l} \sum_{\boldsymbol{s}\in\binom{[n]}{k}} E_l\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right), \tag{27}$$

where the last equality follows by observing that any lth order ($l > k$) minor of $\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*$ is rank deficient, and hence $E_l\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right) = 0$.

Consider an index set $\boldsymbol{r} \in \binom{[m]}{l}$ with $l \leq k$, and denote by $\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right)_{\boldsymbol{r}}$ the submatrix of $\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*$ with rows and columns coming from the index set $\boldsymbol{r}$. One can then verify that

$$\det\left(\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right)_{\boldsymbol{r}}\right) = \det\left(\boldsymbol{B}_{\boldsymbol{r},\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{r},\boldsymbol{s}}^*\right) = \sum_{\tilde{\boldsymbol{r}}\in\binom{\boldsymbol{s}}{l}} \det\left(\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}^*\right),$$

where the last equality arises from the Cauchy-Binet formula (e.g. [36]). Some algebraic manipulation yields that for any $l \leq k$:

$$\sum_{\boldsymbol{s}\in\binom{[n]}{k}} E_l\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right) = \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \sum_{\boldsymbol{r}\in\binom{[m]}{l}} \det\left(\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right)_{\boldsymbol{r}}\right) = \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \sum_{\boldsymbol{r}\in\binom{[m]}{l}} \sum_{\tilde{\boldsymbol{r}}\in\binom{\boldsymbol{s}}{l}} \det\left(\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}^*\right)$$
$$= \sum_{\boldsymbol{r}\in\binom{[m]}{l}} \sum_{\tilde{\boldsymbol{r}}\in\binom{[n]}{l}} \sum_{\boldsymbol{s}:\tilde{\boldsymbol{r}}\subseteq\boldsymbol{s}} \det\left(\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}^*\right)$$
$$\overset{(a)}{=} \sum_{\boldsymbol{r}\in\binom{[m]}{l}} \sum_{\tilde{\boldsymbol{r}}\in\binom{[n]}{l}} \binom{n-l}{k-l} \det\left(\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}^*\right)$$
$$\overset{(b)}{=} \sum_{\boldsymbol{r}\in\binom{[m]}{l}} \binom{n-l}{k-l} = \binom{n-l}{k-l}\binom{m}{l}, \tag{28}$$

where (a) follows from the fact that the number of k-combinations (out of $[n]$) containing $\tilde{\boldsymbol{r}}$ (an l-combination) is $\binom{n-l}{k-l}$, and (b) follows from the Cauchy-Binet formula and the fact that $\boldsymbol{B}\boldsymbol{B}^* = \boldsymbol{I}_m$, i.e.

$$\sum_{\tilde{\boldsymbol{r}}\in\binom{[n]}{l}} \det\left(\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}\boldsymbol{B}_{\boldsymbol{r},\tilde{\boldsymbol{r}}}^*\right) = \det\left(\boldsymbol{B}_{\boldsymbol{r},[n]}\boldsymbol{B}_{\boldsymbol{r},[n]}^*\right) = \det\left(\boldsymbol{I}_l\right) = 1.$$

Substituting (28) into (27) yields

$$\sum_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I}_m + \boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right) = \epsilon^m \binom{n}{k} + \sum_{l=1}^{k} \epsilon^{m-l} \sum_{\boldsymbol{s}\in\binom{[n]}{k}} E_l\left(\boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right)$$
$$= \sum_{l=0}^{k} \binom{n-l}{k-l}\binom{m}{l} \epsilon^{m-l}.$$

By observing that $\boldsymbol{B}_{\boldsymbol{s}}$ is a tall $m \times k$ matrix, one has

$$\det\left(\epsilon \boldsymbol{I}_k + \boldsymbol{B}_{\boldsymbol{s}}^*\boldsymbol{B}_{\boldsymbol{s}}\right) = \epsilon^{k-m} \det\left(\epsilon \boldsymbol{I}_k + \boldsymbol{B}_{\boldsymbol{s}}\boldsymbol{B}_{\boldsymbol{s}}^*\right)$$

$$\Rightarrow \quad \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I}_k + \boldsymbol{B}_{\boldsymbol{s}}^* \boldsymbol{B}_{\boldsymbol{s}}\right) = \sum_{l=0}^{k} \binom{n-l}{k-l}\binom{m}{l}\epsilon^{k-l}.$$

The above expression allows us to derive a crude bound as

$$\begin{aligned}
\binom{n}{k} \min_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I} + \boldsymbol{B}_{\boldsymbol{s}}^* \boldsymbol{B}_{\boldsymbol{s}}\right) \le \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I} + \boldsymbol{B}_{\boldsymbol{s}}^* \boldsymbol{B}_{\boldsymbol{s}}\right) &= \sum_{l=0}^{k} \binom{n-l}{k-l}\binom{m}{l}\epsilon^{k-l} \\
&\le \sum_{l=0}^{k} \binom{n}{k-l}\binom{m}{l}\epsilon^{k-l} = \sum_{l=0}^{k} \binom{n}{l}\binom{m}{m-k+l}\epsilon^{l} \\
&= \binom{m}{k} \sum_{l=0}^{k} \binom{n}{l} \frac{\binom{m}{m-k+l}}{\binom{m}{k}} \epsilon^{l}.
\end{aligned}$$

Since the term $\binom{m}{m-k+l}/\binom{m}{k}$ can be bounded above as

$$\frac{\binom{m}{m-k+l}}{\binom{m}{k}} = \frac{\frac{m!}{(m-k+l)!(k-l)!}}{\frac{m!}{k!(m-k)!}} = \frac{k!}{(k-l)!\frac{(m-k+l)!}{(m-k)!}} = \frac{\binom{k}{l}}{\binom{m-k+l}{l}} \le \binom{k}{l},$$

we obtain

$$\begin{aligned}
\binom{n}{k} \min_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I} + \boldsymbol{B}_{\boldsymbol{s}}^* \boldsymbol{B}_{\boldsymbol{s}}\right) \le \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I} + \boldsymbol{B}_{\boldsymbol{s}}^* \boldsymbol{B}_{\boldsymbol{s}}\right) &\le \binom{m}{k} \sum_{l=0}^{k} \binom{n}{l} \frac{\binom{m}{m-k+l}}{\binom{m}{k}} \epsilon^{l} \\
&\le \binom{m}{k} \sum_{l=0}^{k} \binom{n}{l}\binom{k}{l} \epsilon^{l} \\
&\le \binom{m}{k} \sum_{l=0}^{k} \binom{n+k}{2l} \left(\sqrt{\epsilon}\right)^{2l} \\
&\le \binom{m}{k} \sum_{i=0}^{n+k} \binom{n+k}{i} \left(\sqrt{\epsilon}\right)^{i} \\
&= \binom{m}{k} \left(1+\sqrt{\epsilon}\right)^{n+k},
\end{aligned} \tag{29}$$

where the last equality follows from the binomial theorem.

(2) Since $\boldsymbol{B}_i$ has orthonormal rows, applying the inequality of arithmetic and geometric means yields

$$\begin{aligned}
\sum_{\boldsymbol{s}\in\binom{[n]}{k}} \left(\prod_{i=1}^{p} \det\left(\epsilon \boldsymbol{I}_k + \left(\boldsymbol{B}_i\right)_{\boldsymbol{s}}^* \left(\boldsymbol{B}_i\right)_{\boldsymbol{s}}\right)\right)^{\frac{1}{p}} &\le \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \frac{\sum_{i=1}^{p} \det\left(\epsilon \boldsymbol{I}_k + \left(\boldsymbol{B}_i\right)_{\boldsymbol{s}}^* \left(\boldsymbol{B}_i\right)_{\boldsymbol{s}}\right)}{p} \\
&= \frac{1}{p} \sum_{i=1}^{p} \left\{ \sum_{\boldsymbol{s}\in\binom{[n]}{k}} \det\left(\epsilon \boldsymbol{I}_k + \left(\boldsymbol{B}_i\right)_{\boldsymbol{s}}^* \left(\boldsymbol{B}_i\right)_{\boldsymbol{s}}\right) \right\} \\
&\le \binom{m}{k} \left(1+\sqrt{\epsilon}\right)^{n+k},
\end{aligned}$$

where the second inequality follows from (29). Since $\boldsymbol{M}_\alpha$ has orthonormal rows, applying (29) yields

$$\binom{n}{k}\min_{\boldsymbol{s}\in\binom{[n]}{k}}\left(\prod_{i=1}^{p}\det\left(\epsilon\boldsymbol{I}_k+(\boldsymbol{B}_i)_{\boldsymbol{s}}^{*}(\boldsymbol{B}_i)_{\boldsymbol{s}}\right)\right)^{\frac{1}{p}}\leq\sum_{\boldsymbol{s}\in\binom{[n]}{k}}\left(\prod_{i=1}^{p}\det\left(\epsilon\boldsymbol{I}_k+(\boldsymbol{B}_i)_{\boldsymbol{s}}^{*}(\boldsymbol{B}_i)_{\boldsymbol{s}}\right)\right)^{\frac{1}{p}}$$
$$\leq\binom{m}{k}\left(1+\sqrt{\epsilon}\right)^{n+k}. \tag{30}$$

Therefore,

$$\begin{aligned}\min_{\boldsymbol{s}\in\binom{[n]}{k}}\frac{1}{np}\sum_{i=1}^{p}\log\det\left(\epsilon\boldsymbol{I}_k+(\boldsymbol{B}_i)_{\boldsymbol{s}}^{*}(\boldsymbol{B}_i)_{\boldsymbol{s}}\right)&=\min_{\boldsymbol{s}\in\binom{[n]}{k}}\frac{1}{n}\log\left(\prod_{i=1}^{p}\det\left(\epsilon\boldsymbol{I}_k+(\boldsymbol{B}_i)_{\boldsymbol{s}}^{*}(\boldsymbol{B}_i)_{\boldsymbol{s}}\right)\right)^{\frac{1}{p}}\\&\leq\frac{1}{n}\log\left(\frac{\binom{m}{k}}{\binom{n}{k}}\left(1+\sqrt{\epsilon}\right)^{n+k}\right)\\&=\frac{1}{n}\log\binom{m}{k}-\frac{1}{n}\log\binom{n}{k}+\frac{n+k}{n}\log\left(1+\sqrt{\epsilon}\right)\\&\leq\frac{1}{n}\log\binom{m}{k}-\frac{1}{n}\log\binom{n}{k}+2\sqrt{\epsilon}.\end{aligned}$$

When (n,k,m) are all large numbers, the entropy approximation (26) allows us to approximate the above bound as

$$\begin{aligned}\min_{\boldsymbol{s}\in\binom{[n]}{k}}\frac{1}{np}\sum_{i=1}^{p}\log\det\left(\epsilon\boldsymbol{I}_k+(\boldsymbol{B}_i)_{\boldsymbol{s}}^{*}(\boldsymbol{B}_i)_{\boldsymbol{s}}\right)&\leq\frac{m}{n}\mathcal{H}\left(\frac{k}{m}\right)-\mathcal{H}\left(\frac{k}{n}\right)+\frac{\log(n+1)}{n}+2\sqrt{\epsilon}\\&=\alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right)-\mathcal{H}(\beta)+2\sqrt{\epsilon}+\frac{\log(n+1)}{n}.\end{aligned}\tag{31}$$

APPENDIX C
PROOF OF THEOREM 8

We focus on real-valued random matrices in this theorem. The quantity of interest can be further lower bounded by

$$\begin{aligned}\log\det\left(\epsilon\boldsymbol{I}_k+\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)&=\log\det\left(\epsilon\boldsymbol{M}\boldsymbol{M}^T+\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)-\log\det\left(\boldsymbol{M}\boldsymbol{M}^T\right)\\&\geq\log\det\left(\frac{\epsilon}{k}\sigma_{\min}\left(\boldsymbol{M}\boldsymbol{M}^T\right)\boldsymbol{I}_k+\frac{1}{k}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)-\log\det\left(\frac{1}{k}\boldsymbol{M}\boldsymbol{M}^T\right),\end{aligned}$$

which can hopefully separate $\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T$ and $\boldsymbol{M}\boldsymbol{M}^T$ if $\sigma_{\min}\left(\boldsymbol{M}\boldsymbol{M}^T\right)$ is a constant bounded away from zero. The concentration of the least singular value of a rectangular random matrix with independent sub-Gaussian entries has been largely studied in the random matrix literature (e.g. [28], [40]), which we cite as follows.

Lemma 1. *Suppose that $m<(1-\delta)n$ for some constant $\delta>0$. Let $\boldsymbol{M}$ be an $m\times n$ real-valued random matrix whose entries are jointly independent sub-Gaussian random variables with zero mean and unit variance. Then there exist constants $C,c>0$ such that*

$$\sigma_{\min}\left(\boldsymbol{M}\boldsymbol{M}^{*}\right)>Cm$$

with probability at least $1-\exp(-cn)$. *In particular, if the entries of* $\boldsymbol{M}$ *are i.i.d. Gaussian random variables with zero mean and unit variance, then*

$$\sigma_{\min}\left(\boldsymbol{M}\boldsymbol{M}^{*}\right) > \frac{1}{2}\left(\sqrt{n}-\sqrt{m}\right)^{2}$$

with probability at least $1-4\exp(-cn)$.

Setting $m=k$, we can derive that with probability exceeding $1-4\exp(-cn)$,

$$\log\det\left(\epsilon\boldsymbol{I}_k+\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)\geq\log\det\left(\epsilon C\boldsymbol{I}_k+\frac{1}{k}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)-\log\det\left(\frac{1}{k}\boldsymbol{M}\boldsymbol{M}^T\right) \tag{32}$$

holds for general independent sub-Gaussian matrices, and

$$\log\det\left(\epsilon\boldsymbol{I}_k+\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)\geq\log\det\left(\epsilon\frac{\left(1-\sqrt{\beta}\right)^2}{2\beta}\boldsymbol{I}_k+\frac{1}{k}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)-\log\det\left(\frac{1}{k}\boldsymbol{M}\boldsymbol{M}^T\right) \tag{33}$$

holds for i.i.d. Gaussian matrices.

The next step is to quantify the term $\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)$ for some small $\epsilon>0$. The following lemma characterizes the concentration of measure with respect to this term.

Lemma 2. *Suppose that* $\frac{k}{p}\in(0,1]$ *is a fixed constant. Consider a real-valued random matrix* $\boldsymbol{A}=[\zeta_{ij}]_{1\leq i\leq k,1\leq j\leq p}$ *where* ζ_{ij} *are jointly independent with zero mean and unit variance, and satisfy one of the following conditions:*

(a) ζ_{ij} *is almost surely bounded by a constant* D*;*

(b) ζ_{ij} *satisfies the logarithmic Sobolev inequality with uniformly bounded constant* c_{LS}*.*

Then, for any $\epsilon>0$ *and* $\delta>\frac{4D\sqrt{\pi}}{\epsilon\sqrt{k}(k+p)}$, *there exists a constant* $\tilde{c}>0$ *such that*

$$\mathbb{P}\left(\left|\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)-\mathbb{E}\left(\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)\right|>\delta\right)\leq4\exp\left(-\tilde{c}\epsilon^2\delta^2k^3\right), \tag{34}$$

and

$$\mathbb{P}\left(\left|\frac{1}{p}\log\det{}^{\epsilon}\left(\frac{1}{p}\boldsymbol{A}\boldsymbol{A}^T\right)-\mathbb{E}\left(\frac{1}{p}\log\det{}^{\epsilon}\left(\frac{1}{p}\boldsymbol{A}\boldsymbol{A}^T\right)\right)\right|>\delta\right)\leq4\exp\left(-\tilde{c}\epsilon^2\delta^2p^3\right). \tag{35}$$

Proof: Observe that the Lipschitz constant of $\log(\epsilon+x)$ is upper bounded by $1/\epsilon$ when $x\geq0$. If ζ_{ij} is almost surely bounded by D and hence each entry of $\frac{1}{\sqrt{k}}\boldsymbol{A}$ is bounded by $\frac{D}{\sqrt{k}}$, then applying [29, Corollary 1.8(a)] leads to

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)-\mathbb{E}\left(\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)\right|>\frac{k+p}{k}\delta\right)\\&\leq4\exp\left(-\frac{\epsilon^2}{4D^2}\left(\delta-\frac{2D\sqrt{\pi}}{\epsilon\sqrt{k}\left(k+p\right)}\right)^2k\left(k+p\right)^2\right).\end{aligned}$$

Setting δ to be a positive constant independent of k, we have for sufficiently large k that

$$\mathbb{P}\left(\left|\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)-\mathbb{E}\left(\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)\right|>\frac{k+p}{k}\delta\right)\leq4\exp\left(-\frac{\epsilon^2\delta^2}{4D^2}k^3\right).$$

If ζ_{ij} satisfies the logarithmic Sobolev inequality with uniformly bounded constant c_{LS}, then the logarithmic Sobolev constant is bounded above by $\frac{1}{k}c_{\text{LS}}$, and hence [29, Corollary 1.8(b)] leads to

$$\mathbb{P}\left(\left|\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)-\mathbb{E}\left(\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)\right|>\frac{k+p}{k}\delta\right)\leq2\exp\left(-\frac{\epsilon^2\delta^2k\left(k+p\right)^2}{2c_{\text{LS}}}\right).$$

The proof is complete by observing that k/p is a given constant.

Given that the Lipschitz constant of the function $\log^{\epsilon} x$ is also $1/\epsilon$, the concentration result for $\frac{1}{p}\log\det^{\epsilon}\left(\frac{1}{p}\boldsymbol{A}\boldsymbol{A}^T\right)$ follows with the same machinery. ■

Now that we have established the concentration results for $\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)$, it remains to determine $\mathbb{E}\left(\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)$. This is established in the following lemma.

Lemma 3. *Let matrix $\boldsymbol{A}=(\zeta_{ij})_{1\leq i,j\leq k}$ be a real-valued random matrix such that all entries are jointly independent with zero mean and unit variance. For any small constant $\epsilon>0$, we have*

$$\mathbb{E}\left[\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right]\leq\frac{1}{k}\log\mathbb{E}\left[\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right]=-1+O\left(\frac{\log k}{k}\right)+2\sqrt{\epsilon}. \tag{36}$$

Additionally, under Condition (a) or (b) of Lemma 2, $\boldsymbol{A}$ satisfies

$$\mathbb{E}\left(\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)\geq-1-O\left(\frac{\log\epsilon}{k}\right). \tag{37}$$

Proof: See Appendix E. ■

Let us fix $\delta=\frac{\log k}{k}$. Combining Lemma 3 with Lemma 2 yields that under the assumptions of Lemma 2, one has for any small constant $\epsilon>0$, for sufficiently large k we have

$$O\left(\left|\frac{\log\epsilon}{k}\right|\right)\leq\frac{\log k}{k}$$

and hence

$$\frac{1}{k}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\in\left[-1-\frac{2\log k}{k},-1+O\left(\frac{\log k}{k}\right)+2\sqrt{\epsilon}\right] \tag{38}$$

holds with probability exceeding $1-4\exp\left(-\tilde{c}\epsilon^2k\log k\right)$ for some constant $\tilde{c}>0$.

By our assumption, $\boldsymbol{M}_{\boldsymbol{s}}$ satisfies the conditions of Lemma 2. For any alphabet $\mathcal{S}\subseteq\binom{[n]}{k}$, it contains at most $\binom{n}{k}\approx e^{n\mathcal{H}(\beta)}$ different states. Hence, applying the union bound gives

$$\forall\mathbf{s}:\ \frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)\in\left[-\beta-\frac{2\log k}{n},-\beta+O\left(\frac{\log k}{k}\right)+2\beta\sqrt{\epsilon}\right] \tag{39}$$

with probability exceeding $1-4\exp\left(n\mathcal{H}\left(\beta\right)\right)\exp\left(-\tilde{c}\epsilon^2k\log k\right)\geq1-4\exp\left(-\hat{c}\epsilon^2k\log k\right)$. Here, $\hat{c}$ is some positive constant.

The last step is to quantify $\frac{1}{n}\log\det\left(\frac{1}{k}\boldsymbol{M}\boldsymbol{M}^T\right)$. This is evaluated in the following lemma.

Lemma 4. *Suppose that $\alpha:=\frac{m}{n}$ is a fixed constant independent from (m,n) and that there exists a constant $\delta_{\mathrm{g}}>0$ such that $\alpha\leq1-\delta_{\mathrm{g}}$. Let matrix $\boldsymbol{A}=(\zeta_{ij})_{1\leq i\leq m,1\leq j\leq n}$ be a real-valued random matrix such that all entries are jointly independent with zero mean and unit variance.*

(1) Under Condition (a) or (b) of Lemma 2, for any $\epsilon<\frac{1}{e}\left(1-\sqrt{\alpha}\right)^2$, there exists a constant $c_7>0$ such that

$$\frac{1}{n}\mathbb{E}\log\det^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)\leq(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+\frac{3\log m}{n}, \tag{40}$$

for sufficiently large n, and

$$\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)\leq\frac{1}{n}\log\det^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)\leq(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+\frac{4\log m}{n}$$

holds with probability exceeding $1-4\exp\left(-c_7\epsilon^2 n\log^2 m\right)$.

(2) If $\boldsymbol{A}$ *is drawn from i.i.d. Gaussian ensemble, then*

$$\frac{1}{n}\log \textit{det}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)\geq(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+O\left(\frac{\log^2 n}{\sqrt{n}}\right)$$

with probability exceeding $1-\exp\left(c_9 n\log n\right)$ *for some constant* $c_9>0$.

Proof: See Appendix F. ■

Since $\boldsymbol{M}$ satisfies the assumptions of Lemma 4 with $\beta=\alpha$, simple manipulation gives

$$\begin{aligned}\frac{1}{n}\log\det\left(\frac{1}{k}\boldsymbol{M}\boldsymbol{M}^T\right)&=\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right)+\frac{k}{n}\log\frac{n}{k}\\&\leq(1-\beta)\log\frac{1}{1-\beta}-\beta+\beta\log\frac{1}{\beta}+\frac{4\log k}{n}\end{aligned}\tag{41}$$

with probability exceeding $1-5\exp\left(-\hat{c}n\log^2 k\right)$.

The above results (32), (33), (39) and (41) taken collectively yield the following, under the condition of Theorem 8, one has

$$\forall\boldsymbol{s}:\quad\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)\geq-\mathcal{H}\left(\beta\right)-\frac{5\log k}{n}\tag{42}$$

with probability exceeding $1-\exp\left(-cn\right)$ for some $c>0$. Combining this lower bound with the upper bound developed in Theorem 7 (with $\alpha=\beta$) completes our proof.

Appendix D
Proof of Theorem 9

Our goal is to evaluate $\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{M}_{\boldsymbol{s}}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)$ for some small $\epsilon>0$. We first define two Wishart matrices $\Xi_{\backslash\boldsymbol{s}}:=\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T-\frac{1}{n}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T$ and $\Xi_{\boldsymbol{s}}:=\frac{1}{n}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T$. Apparently, $\Xi_{\boldsymbol{s}}\sim\mathcal{W}_m\left(k,\frac{1}{n}\boldsymbol{I}_m\right)$ and $\Xi_{\backslash\boldsymbol{s}}\sim\mathcal{W}_m\left(n-k,\frac{1}{n}\boldsymbol{I}_m\right)$. When $1-\alpha>\beta$, i.e. $n-k>m$, the Wishart matrix $\Xi_{\backslash\boldsymbol{s}}$ is invertible with probability 1.

One difficulty in evaluating $\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{M}_{\boldsymbol{s}}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)$ is that $\boldsymbol{M}_{\boldsymbol{s}}$ and $\boldsymbol{M}\boldsymbol{M}^T$ are not independent. This motivates us to decouple them first as follows

$$\begin{aligned}\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{M}_{\boldsymbol{s}}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)&=\epsilon^{k-m}\det\left(\epsilon\boldsymbol{I}_m+\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\frac{1}{n}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)\\&=\epsilon^{k-m}\det\left(\epsilon\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T+\frac{1}{n}\boldsymbol{M}_{\boldsymbol{s}}\boldsymbol{M}_{\boldsymbol{s}}^T\right)\det\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\\&=\epsilon^{k-m}\det\left(\epsilon\Xi_{\backslash\boldsymbol{s}}+(1+\epsilon)\,\Xi_{\boldsymbol{s}}\right)\det\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\\&=\epsilon^{k-m}\det\left(\epsilon\boldsymbol{I}_m+(1+\epsilon)\,\Xi_{\boldsymbol{s}}\Xi_{\backslash\boldsymbol{s}}^{-1}\right)\det\left(\Xi_{\backslash\boldsymbol{s}}\right)\det\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\\&=\det\left(\epsilon\boldsymbol{I}_k+(1+\epsilon)\,\frac{1}{n}\boldsymbol{M}_{\boldsymbol{s}}^T\Xi_{\backslash\boldsymbol{s}}^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)\det\left(\Xi_{\backslash\boldsymbol{s}}\right)\det\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\end{aligned}$$

or, equivalently,

$$\begin{aligned}
&\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{M}_{\boldsymbol{s}}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)\\
=&\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+(1+\epsilon)\,\frac{1}{n}\boldsymbol{M}_{\boldsymbol{s}}^T\Xi_{\backslash\boldsymbol{s}}^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)+\frac{1}{n}\log\det\left(\Xi_{\backslash\boldsymbol{s}}\right)-\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right). \qquad (43)
\end{aligned}$$

The point of developing this identity (43) is to decouple the left-hand side of (43) through 3 matrices $\boldsymbol{M}_{\boldsymbol{s}}^T\Xi_{\backslash\boldsymbol{s}}^{-1}\boldsymbol{M}_{\boldsymbol{s}}$, $\Xi_{\backslash\boldsymbol{s}}$ and $\boldsymbol{M}\boldsymbol{M}^T$. In particular, since $\boldsymbol{M}_{\boldsymbol{s}}$ and $\Xi_{\backslash\boldsymbol{s}}$ are jointly independent, we can examine the concentration of measure for $\boldsymbol{M}_{\boldsymbol{s}}$ and $\Xi_{\backslash\boldsymbol{s}}$ separately when evaluating $\boldsymbol{M}_{\boldsymbol{s}}^T\Xi_{\backslash\boldsymbol{s}}^{-1}\boldsymbol{M}_{\boldsymbol{s}}$.

The second and third terms of (43) can be evaluated through Lemma 4, which indicates that

$$\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^T\right)\leq-(1-\alpha)\log(1-\alpha)-\alpha+\frac{4\log m}{n} \qquad (44)$$

and for all $\boldsymbol{s}\in\binom{[n]}{k}$:

$$\begin{aligned}
\frac{1}{n}\log\det\left(\Xi_{\backslash\boldsymbol{s}}\right)=&\frac{n-k}{n}\frac{1}{n-k}\log\det\left(\frac{n}{n-k}\Xi_{\backslash\boldsymbol{s}}\right)+\frac{1}{n}\log\det\left(\frac{n-k}{n}\boldsymbol{I}\right)\\
\geq&(1-\beta)\left\{-\left(1-\frac{\alpha}{1-\beta}\right)\log\left(1-\frac{\alpha}{1-\beta}\right)-\frac{\alpha}{1-\beta}\right\}+\alpha\log(1-\beta)+O\left(\frac{\log^2 n}{n^{1/2}}\right)\\
\geq&-(1-\alpha-\beta)\log\left(1-\frac{\alpha}{1-\beta}\right)-\alpha+\alpha\log(1-\beta)+O\left(\frac{\log^2 n}{n^{1/2}}\right). \qquad (45)
\end{aligned}$$

hold simultaneously with probability exceeding $1-\exp(-c_9 n)$.

Our main task is then to quantify $\log\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{M}_{\boldsymbol{s}}^T\Xi_{\backslash\boldsymbol{s}}^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)$, which we derive through the following lemma.

Lemma 5. *Suppose that $m>k$. Let matrix $\boldsymbol{A}=(\zeta_{ij})_{1\leq i\leq m,1\leq j\leq k}$ and $\boldsymbol{B}=(\xi_{ij})_{1\leq i,j\leq m}$ be two independent matrix ensembles such that $\zeta_{ij}\sim\mathcal{N}(0,1)$ are jointly independent, and $\boldsymbol{B}\sim\mathcal{W}_m(n-k,\boldsymbol{I}_m)$. Then*

$$\begin{aligned}
&\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{A}\right)\\
\geq&-(\alpha-\beta)\log(\alpha-\beta)+\alpha\log\alpha+(1-\alpha-\beta)\log\left(1-\frac{\beta}{1-\alpha}\right)-\beta\log(1-\alpha)+O\left(\frac{\log n}{\sqrt{n}}\right)
\end{aligned}$$

holds with probability exceeding $1-2\exp(-\tilde{c}_8 n\log n)$ for some constant c_8.

Proof: See Appendix G. ■

Lemma 5 develops a lower bound for $\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{M}_{\boldsymbol{s}}^T\Xi_{\backslash\boldsymbol{s}}^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)$. This together with (44), (45) and (43) yields that

$$\begin{aligned}
&\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{M}_{\boldsymbol{s}}^T\left(\boldsymbol{M}\boldsymbol{M}^T\right)^{-1}\boldsymbol{M}_{\boldsymbol{s}}\right)\\
\geq&-(\alpha-\beta)\log(\alpha-\beta)+\alpha\log\alpha+(1-\alpha-\beta)\log\left(1-\frac{\beta}{1-\alpha}\right)-\beta\log(1-\alpha)\\
&-(1-\alpha-\beta)\log\left(1-\frac{\alpha}{1-\beta}\right)-\alpha+\alpha\log(1-\beta)+(1-\alpha)\log(1-\alpha)+\alpha+O\left(\frac{\log n}{\sqrt{n}}\right)\\
=&\alpha\mathcal{H}\left(\frac{\beta}{\alpha}\right)-\mathcal{H}(\beta)+O\left(\frac{\log n}{\sqrt{n}}\right)
\end{aligned}$$

with probability exceeding $1-4\exp(-\tilde{c}_8 n\log n)$.

Since there are at most $\binom{n}{k} \approx e^{n\mathcal{H}(\beta)}$ different states $\boldsymbol{s}$, applying a union bound over all states completes the proof.

APPENDIX E
PROOF OF LEMMA 3

(1) The identity (25) gives

$$\mathbb{E}\left(\det\left(\epsilon k\boldsymbol{I}+\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)=(\epsilon k)^{k}+\sum_{i=1}^{k}(\epsilon k)^{k-i}\,\mathbb{E}\left(E_{i}\left(\boldsymbol{A}\boldsymbol{A}^{T}\right)\right),$$

where $E_i\left(\boldsymbol{A}\boldsymbol{A}^T\right)$ denotes the sum of determinants of all i-by-i principal minors of $\boldsymbol{A}\boldsymbol{A}^T$. There is a well known fact that for any $\boldsymbol{G}=(\zeta_{ij})_{1\le i,j\le l}$ with independent entries having mean zero and variance one, one has $\mathbb{E}\det\left(\boldsymbol{G}\boldsymbol{G}^T\right)=l!$. In fact, if we denote by $\prod_l$ the permutation group of l elements, then the Leibniz formula for the determinant gives

$$\det(\boldsymbol{G})=\sum_{\sigma\in\prod_{l}}\mathrm{sgn}(\sigma)\prod_{i=1}^{k}\zeta_{i,\sigma(i)}.$$

Since ζ_{ij} are jointly independent, we have

$$\mathbb{E}\mathrm{det}\left(\boldsymbol{G}\boldsymbol{G}^{T}\right)=\mathbb{E}\left(\det(\boldsymbol{G})\right)^{2}=\sum_{\sigma\in\prod_{l}}\mathbb{E}\prod_{i=1}^{l}\left|\zeta_{i,\sigma(i)}\right|^{2}=\sum_{\sigma\in\prod_{l}}\prod_{i=1}^{l}\mathbb{E}\left|\zeta_{i,\sigma(i)}\right|^{2}=l!. \tag{46}$$

Denote by $\left(\boldsymbol{A}\boldsymbol{A}^T\right)_{\boldsymbol{s}}$ the principal minor of $\boldsymbol{A}\boldsymbol{A}^T$ coming from the index set $\boldsymbol{s}$, then $E_i\left(\boldsymbol{A}\boldsymbol{A}^T\right)$ can be computed as

$$\begin{aligned}\mathbb{E}\left(E_{i}\left(\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)&=\sum_{\boldsymbol{s}\in\binom{[k]}{i}}\mathbb{E}\left(\det\left(\left(\boldsymbol{A}\boldsymbol{A}^{T}\right)_{\boldsymbol{s}}\right)\right)=\sum_{\boldsymbol{s}\in\binom{[k]}{i}}\mathbb{E}\left(\det\left(\boldsymbol{A}_{\boldsymbol{s},[n]}\boldsymbol{A}_{\boldsymbol{s},[n]}^{T}\right)\right)\\&\overset{(a)}{=}\sum_{\boldsymbol{s}\in\binom{[k]}{i}}\sum_{\boldsymbol{r}\in\binom{[k]}{i}}\mathbb{E}\left(\det\left(\boldsymbol{A}_{\boldsymbol{s},\boldsymbol{r}}\boldsymbol{A}_{\boldsymbol{s},\boldsymbol{r}}^{T}\right)\right)\\&\overset{(b)}{=}\binom{k}{i}\binom{k}{i}i!,\end{aligned}$$

where (a) follows from the Cauchy-Binet formula and (b) follows from (46). Combining the above results with simple algebraic manipulation yields

$$\begin{aligned}\mathbb{E}\left(\det\left(\epsilon k\boldsymbol{I}+\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)&=(\epsilon k)^{k}+\sum_{i=1}^{k}(\epsilon k)^{k-i}\binom{k}{i}\frac{k!}{(k-i)!}=k!\sum_{i=0}^{k}\binom{k}{k-i}\frac{(\epsilon k)^{k-i}}{(k-i)!}\\&=k!\sum_{i=0}^{k}\binom{k}{i}\frac{\epsilon^{i}k^{i}}{i!}.\end{aligned} \tag{47}$$

In order to bound this sum, one way is to first identify the largest term. Using the short hand notation $f_\epsilon(j):=\binom{k}{j}\frac{k^j\epsilon^j}{j!}$ and $r_\epsilon(j+1):=\frac{f_\epsilon(j+1)}{f_\epsilon(j)}$, we can obtain

$$r_{\epsilon}(j)=\frac{f_{\epsilon}(j+1)}{f_{\epsilon}(j)}=\frac{\frac{k!}{(k-j-1)!(j+1)!}\frac{k^{j+1}\epsilon^{j+1}}{(j+1)!}}{\frac{k!}{(k-j)!j!}k^{j}\epsilon^{j}j!}=\frac{(k-j)\,k\epsilon}{(j+1)^{2}}.$$

Apparently, $r_\epsilon(j)$ is a decreasing function of j. By setting $r_\epsilon(j^*)=1$, we can derive

$$\frac{(k-j^*)\,k\epsilon}{(j^*+1)^2}=1 \quad \Rightarrow \quad j^*=\frac{-k\epsilon-2+\sqrt{4k^2\epsilon-4+(k\epsilon+2)^2}}{2}$$

which can be bounded by

$$k\left(\sqrt{\epsilon}-\epsilon\right)<\frac{-k\epsilon-2+\sqrt{4k^2\epsilon}}{2}<j^*<\frac{-k\epsilon-2+\sqrt{4k^2\epsilon}+\sqrt{(k\epsilon+2)^2}}{2}=k\sqrt{\epsilon}$$

for large k and small ϵ. Suppose that $j^*=\tilde{\epsilon}k$ for some $\tilde{\epsilon}\in[\sqrt{\epsilon}-\epsilon,\sqrt{\epsilon}]$, then by Stirling's formula

$$\begin{aligned}\frac{1}{k}\log f(j)\le\frac{1}{k}\log f\left(j^*\right)&=\frac{1}{k}\log\left\{\binom{k}{j^*}\frac{k^{j^*}\epsilon^{j^*}}{j^*!}\right\}\\&=-\tilde{\epsilon}\log\tilde{\epsilon}-(1-\tilde{\epsilon})\log\left(1-\tilde{\epsilon}\right)+\tilde{\epsilon}\log k+\tilde{\epsilon}\log\epsilon-\tilde{\epsilon}\log\left(\tilde{\epsilon}k\right)+\tilde{\epsilon}+O\left(\frac{\log k}{k}\right)\\&\le 2\sqrt{\epsilon}+O\left(\frac{\log k}{k}\right)=2\sqrt{\epsilon}+O\left(\frac{\log k}{k}\right)\end{aligned}$$

where the last equality follows from L' Hospital's rule. This together with (47) yields

$$\begin{aligned}\frac{1}{k}\log\mathbb{E}\left(\det\left(\epsilon k\boldsymbol{I}+\boldsymbol{A}\boldsymbol{A}^T\right)\right)&\le\frac{1}{k}\log\left[k!\left(kf(j^*)\right)\right]\\&=\frac{\log\left(k!\right)+\log k}{k}+2\sqrt{\epsilon}+O\left(\frac{\log k}{k}\right)\\&=\log k-1+2\sqrt{\epsilon}+O\left(\frac{\log k}{k}\right)\end{aligned}$$

Together with Jensen's inequality, one has

$$\begin{aligned}\frac{1}{k}\mathbb{E}\left(\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)&=-\log k+\frac{1}{k}\mathbb{E}\left(\log\det\left(\epsilon k\boldsymbol{I}+\boldsymbol{A}\boldsymbol{A}^T\right)\right)\\&\le-\log k+\frac{1}{k}\log\mathbb{E}\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\\&=-1+2\sqrt{\epsilon}+O\left(\frac{\log k}{k}\right).\end{aligned}$$

(2) The lower bound follows from concentration inequality. Define $Y:=\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)-\mathbb{E}\left(\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\right)$, Lemma 2 implies that $\mathbb{P}\left(|Y|>\delta\right)\le 4\exp\left(-\tilde{c}\epsilon^2\delta^2k\right)$. Denote by $f_Y(\cdot)$ the probability density function of Y, we can obtain

$$\begin{aligned}\mathbb{E}\left(e^Y\right)\le\mathbb{E}\left(e^{|Y|}\right)&=\int_0^\infty e^y f_Y\left(y\right)\mathrm{d}y=-e^y\mathbb{P}\left(Y>y\right)|_0^\infty+\int_0^\infty e^y\mathbb{P}\left(Y>y\right)\mathrm{d}y\\&\le 1+\int_0^\infty 4\exp\left(y-\tilde{c}\epsilon^2ky^2\right)\mathrm{d}y\\&<1+4\sqrt{\frac{\pi}{\tilde{c}\epsilon^2}}\exp\left(\frac{1}{4\tilde{c}\epsilon^2k}\right).\end{aligned}$$

Taking logarithms at both sides and plugging in the expression of Y yields

$$\log\mathbb{E}\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)\le\mathbb{E}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)+\log\left[1+4\sqrt{\frac{\pi}{\tilde{c}\epsilon^2}}\exp\left(\frac{1}{4\tilde{c}\epsilon^2k}\right)\right]. \tag{48}$$

Combining (48) and (46) yields that

$$\begin{aligned}\frac{1}{k}\mathbb{E}\log\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right) &\geq \frac{1}{k}\log\mathbb{E}\det\left(\epsilon\boldsymbol{I}+\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)+\frac{\log\epsilon}{k}-O\left(\frac{1}{k}\right)\\ &\geq \frac{1}{k}\log\mathbb{E}\det\left(\frac{1}{k}\boldsymbol{A}\boldsymbol{A}^T\right)+\frac{\log\epsilon}{k}-O\left(\frac{1}{k}\right)=\frac{1}{k}\log\frac{k!}{k^k}+\frac{\log\epsilon}{k}-O\left(\frac{1}{k}\right)\\ &= -1+\frac{\log\epsilon}{k}-O\left(\frac{1}{k}\right).\end{aligned}$$

APPENDIX F
PROOF OF LEMMA 4

(1) We first develop the upper bound. The Cauchy-Binet formula indicates that

$$\mathbb{E}\left(\det\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right)=\sum_{\boldsymbol{s}\in\binom{[n]}{m}}\mathbb{E}\left(\det\left(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^T\right)\right),$$

where $\boldsymbol{s}$ ranges over all m-combinations of $\{1,\cdots,n\}$, and $\boldsymbol{A}_{\boldsymbol{s}}$ is the $m\times m$ minor of $\boldsymbol{A}$ whose columns are the columns of $\boldsymbol{A}$ at indices from s. It has been shown in (46) that for each jointly independent $m\times m$ ensemble $\boldsymbol{A}_{\boldsymbol{s}}$, the determinant satisfies

$$\mathbb{E}\det\left(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^T\right)=m!,$$

which immediately leads to

$$\mathbb{E}\left(\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)\right)=\frac{1}{n^m}\sum_{\boldsymbol{s}\in\binom{[n]}{m}}\mathbb{E}\left(\det\left(\boldsymbol{A}_{\boldsymbol{s}}\boldsymbol{A}_{\boldsymbol{s}}^T\right)\right)=\frac{m!}{n^m}\binom{n}{m}.$$

Besides, using the entropy formula and the following identity [36, Equation 1.46]

$$\log(m!)\leq\log\left(emm^me^{-m}\right)=(m+1)\log m-m+1,$$

we can obtain

$$\begin{aligned}\frac{1}{n}\log\mathbb{E}\left(\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)\right) &\leq -\frac{m}{n}\log n+\frac{(m+1)\log m}{n}-\frac{m}{n}+\frac{1}{n}+\mathcal{H}\left(\frac{m}{n}\right)\\ &\leq -\frac{m}{n}\log n+\frac{m\log m}{n}-\frac{m}{n}+\frac{2\log m}{n}+\mathcal{H}\left(\frac{m}{n}\right)\\ &=(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+\frac{2\log m}{n}.\end{aligned}$$

Define $Z:=\log\det^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)-\mathbb{E}\left(\log\det^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)\right)$, then Lemma 2 implies that $\mathbb{P}\left(|Z|>\tau\right)\leq 4\exp\left(-\tilde{c}\epsilon^2\tau^2n\right)$. We can now derive

$$\begin{aligned}\mathbb{E}\left(e^Z\right)\leq\mathbb{E}\left(e^{|Z|}\right) &= -e^z\mathbb{P}\left(Z>z\right)|_{z=0}^{\infty}+\int_0^{\infty}e^z\mathbb{P}\left(Z>z\right)\mathrm{d}z\\ &\leq 1+\int_0^{\infty}4\exp\left(z-\tilde{c}\epsilon^2nz^2\right)\mathrm{d}z\\ &< 1+4\sqrt{\frac{\pi}{\tilde{c}\epsilon^2}}\exp\left(\frac{1}{4\tilde{c}\epsilon^2n}\right).\end{aligned}$$

Taking logarithms and plugging in the expression of Z yields

$$\log \mathbb{E}\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) \leq \mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) + \log\left[1+4\sqrt{\frac{\pi}{\tilde{c}\epsilon^{2}}}\exp\left(\frac{1}{4\tilde{c}\epsilon^{2}n}\right)\right], \tag{49}$$

which leads to

$$\begin{aligned}\frac{1}{n}\mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) &\geq \frac{1}{n}\log\mathbb{E}\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) - O\left(\frac{\log\epsilon}{n}\right) \geq \frac{1}{n}\log\mathbb{E}\mathrm{det}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) - O\left(\frac{\log\epsilon}{n}\right)\\ &= (1-\alpha)\log\frac{1}{1-\alpha} - \alpha - O\left(\frac{\log n + \log\epsilon}{n}\right).\end{aligned} \tag{50}$$

On the other hand, setting $\tau = 1$ in the bound $\mathbb{P}\left(|Z| > \tau\right) \leq 4\exp\left(-\tilde{c}\epsilon^{2}\tau^{2}n\right)$ indicates that with probability at least $1-4\exp\left(-\tilde{c}\epsilon^{2}n\right)$, we have

$$|Z| = \left|\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) - \mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right| \leq 1$$

or, equivalently,

$$\frac{1}{e}e^{\mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)} \leq \mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) \leq e\cdot e^{\mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)}.$$

Jensen's inequality implies that

$$e^{\mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)} \leq \mathbb{E}e^{\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)} = \mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right),$$

and hence one also has

$$\frac{1}{e}e^{\mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)} \leq \mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) \leq e\cdot\mathbb{E}\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)$$

with probability exceeding $1-4\exp\left(-\tilde{c}\epsilon^{2}n\right)$.

Besides, the tail distribution for $\sigma_{\min}\left(\boldsymbol{A}\right)$ satisfies [41, Theorem 1.1]

$$\mathbb{P}\left(\sigma_{\min}\left(\frac{1}{\sqrt{n}}\boldsymbol{A}\right) \leq \tilde{\epsilon}\left(1-\sqrt{\alpha}\right)\right) \leq \left(C\tilde{\epsilon}\right)^{(1-\alpha)n} + e^{-cn}$$

for some constants $C, c > 0$. For a small constant $\epsilon < \frac{1}{2e}\left(1-\sqrt{\alpha}\right)^{2}$, taking $\tilde{\epsilon} = \frac{\sqrt{\epsilon}}{1-\sqrt{\alpha}}$ gives

$$\mathbb{P}\left(\sigma_{\min}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) \leq \epsilon\right) \leq \exp\left(-\tilde{c}_{\mathrm{m}}n\right)$$

for some constant $\tilde{c}_{\mathrm{m}} > 0$. This basically implies that

$$\mathbb{P}\left(\mathrm{det}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) = \mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right) \geq 1 - \exp\left(-\tilde{c}_{\mathrm{m}}n\right). \tag{51}$$

Hence, the union bound implies that

$$\mathrm{det}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) = \mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) \quad \text{and} \quad \mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) > e^{-1}e^{\mathbb{E}\left(\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)}$$

hold with probability exceeding $1-2\exp\left(-\breve{c}n\right)$ for some constant $\breve{c}$. Therefore,

$$\mathbb{E}\,\mathrm{det}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) \geq \left(1-2\exp\left(-\breve{c}n\right)\right)e^{-1}e^{\mathbb{E}\left(\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)},$$

which immediately implies that for sufficiently large n,

$$\frac{1}{n}\log\mathbb{E}\,\mathrm{det}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right) \geq \frac{1}{n}\mathbb{E}\left(\log\mathrm{det}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right) - \frac{2}{n}.$$

Equivalently, we have

$$\frac{1}{n}\mathbb{E}\left(\log\det{}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)\leq\frac{1}{n}\log\mathbb{E}\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)+\frac{2}{n}\leq(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+\frac{3\log m}{n}. \tag{52}$$

Setting $\delta=\log m/n$ in Lemma 2 then leads to

$$\mathbb{P}\left\{\frac{1}{n}\log\det{}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\leq\frac{1}{n}\mathbb{E}\left(\log\det{}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)+\frac{\log m}{n}\right\}\geq1-4\exp\left(-\tilde{c}\epsilon^{2}n\log^{2}m\right).$$

Using the upper bound (52) and the fact that $\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\leq\frac{1}{n}\log\det^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)$ immediately gives

$$\mathbb{P}\left\{\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\leq(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+\frac{4\log m}{n}\right\}\geq1-4\exp\left(-\tilde{c}\epsilon^{2}n\log^{2}m\right).$$

(2) In order to derive a lower bound on $\log\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)$, it is helpful to first estimate the number of eigenvalues of $\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}$ that are smaller than ϵ, i.e. $\sum_{i=1}^{n}\mathbf{1}_{[0,\epsilon]}\left(\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)$. Since the indicator function $\mathbf{1}_{[0,\epsilon]}\left(\cdot\right)$ is discontinuous, we define instead a continuous function $g_{\epsilon}(x)$ not smaller than $\mathbf{1}_{[0,\epsilon]}(x)$ such that

$$g_{\epsilon}(x)=\begin{cases}1, & \text{if }0\leq x\leq\epsilon;\\ -x/\epsilon+1, & \text{if }\epsilon<x\leq2\epsilon;\\ 0, & \text{else.}\end{cases}$$

Clearly, $g_{\epsilon}(\cdot)$ has a Lipschitz constant $1/\epsilon$ and satisfies $\mathbf{1}_{[0,\epsilon]}(x)\leq g_{\epsilon}(x)$. For any small constant $0<\epsilon<\frac{1}{2e}$, we have the following crude bound

$$\sum_{i=1}^{n}\mathbf{1}_{[0,\epsilon]}\left(\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)\leq\sum_{i=1}^{n}g_{\epsilon}\left(\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)<\sum_{i:\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\leq2\epsilon}\log\frac{1}{\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)}. \tag{53}$$

Since $x\log\frac{1}{x}$ is an increasing function for $x<2\epsilon<e^{-1}$, we have $\log\frac{1}{x}\leq\frac{2\epsilon\log\frac{1}{2\epsilon}}{x}$ and hence

$$\sum_{i:\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\leq\epsilon}\log\frac{1}{\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)}\leq\sum_{i:\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\leq\epsilon}\frac{-2\epsilon\log\left(2\epsilon\right)}{\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)}\leq2\epsilon\log\frac{1}{2\epsilon}\mathrm{tr}\left(\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)^{-1}\right).$$

This together with (53) yields

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}g_{\epsilon}\left(\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)\right)\leq\left(2\epsilon\log\frac{1}{2\epsilon}\right)\frac{1}{n}\mathbb{E}\mathrm{tr}\left(\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)^{-1}\right)=\frac{2m\epsilon\log\frac{1}{2\epsilon}}{n-m-1}\leq\frac{\alpha}{1-\alpha}3\epsilon\log\frac{1}{2\epsilon},$$

where the equality above follows from the property of Wishart matrices (e.g. [42, Theorem 2.2.8]).

Note that $g_{\epsilon}(x)$ has Lipschitz constant $1/\epsilon$ and standard Gaussian measure has logarithmic Sobolev constant $c_{\mathrm{LS}}=1$. Applying [29, Corollary 1.8(b)] yields that for any $\delta>0$

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}g_{\epsilon}\left(\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)-\frac{\alpha}{1-\alpha}3\epsilon\log\frac{1}{2\epsilon}>\delta\right)\\ \leq&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}g_{\epsilon}\left(\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)-\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}g_{\epsilon}\left(\lambda_{i}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)\right)>\delta\right)\\ \leq&\ 2\exp\left(-2\epsilon^{2}\delta^{2}m^{3}\right).\end{aligned}$$

Since $g(x)$ is an upper bound on $\mathbf{1}_{[0,\epsilon]}(x)$, this implies that

$$\frac{\text{card}\left\{i \mid \lambda_i\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) < \epsilon\right\}}{n} < \frac{3\alpha}{1-\alpha}\epsilon\log\frac{1}{2\epsilon} + \delta \tag{54}$$

with probability exceeding $1 - 2\exp\left(-2\epsilon^2\delta^2 m^3\right)$.

Now that we have an estimate of the number of small eigenvalues, our next task is to estimate the influence by the set of small eigenvalues. It has been shown in [43, Theorem 4.5] that

$$\begin{aligned}\mathbb{P}\left(\frac{\lambda_{\max}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)}{\lambda_{\min}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)} > n^4\right) \leq \mathbb{P}\left(\sqrt{\frac{\lambda_{\max}\left(\boldsymbol{A}\boldsymbol{A}^T\right)}{\lambda_{\min}\left(\boldsymbol{A}\boldsymbol{A}^T\right)}} > n\right) &\leq \frac{1}{\sqrt{2\pi}}\left(\frac{6.414}{n}\right)^{n-m+1} \\ &\leq \exp\left(-c_5 n\log n\right). \end{aligned}\tag{55}$$

for some constant $c_5 > 0$.

Now that we have a tail upper bound for the condition number, we can lower bound $\lambda_{\min}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)$ by developing a lower bound on $\lambda_{\max}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)$. By setting $\epsilon = n^{-1/4}$ and $\delta = n^{-1/4}$ in (54) and hence $\frac{\alpha}{1-\alpha}\epsilon\log\epsilon + \delta \ll 1/2$, we can derive a crude lower bound such that

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\mathbf{A}\mathbf{A}^T\right) < \frac{1}{n^{1/4}}\right) &\leq \mathbb{P}\left(\frac{\text{card}\left\{i \mid \lambda_i\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) < \frac{1}{n^{1/4}}\right\}}{n} > \frac{1}{2}\right) \\ &\leq 2\exp\left(-c_4 m^2\right).\end{aligned}$$

This together with (55) yields

$$\begin{aligned}&\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) \leq \frac{1}{n^{4.25}}\right) \\ \leq&\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) < \frac{1}{n^{1/4}}\right) + \mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) \leq \frac{1}{n^{4.25}} \text{ and } \lambda_{\max}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) > \frac{1}{n^{1/4}}\right) \\ \leq&\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) < \frac{1}{n^{1/4}}\right) + \mathbb{P}\left(\frac{\lambda_{\max}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)}{\lambda_{\min}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)} > n^4\right) \\ \leq&\exp\left(-\tilde{c}_5 n\log n\right)\end{aligned}\tag{56}$$

for some constant $\tilde{c}_5 > 0$.

Set $\epsilon = n^{-1/2}$, $\delta = n^{-1/2}\log n$ from now on, which indicates that

$$\frac{\text{card}\left\{i : \lambda_i\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) < \epsilon\right\}}{n} = O\left(\frac{\log n}{\sqrt{n}}\right) \tag{57}$$

with probability at least $1 - \exp\left(-\hat{c}_6 n\log^2 n\right)$ for some constant $\hat{c}_6 > 0$. Note that

$$\begin{aligned}0 \leq \frac{1}{n}\log\det{}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) - \frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) &= \frac{1}{n}\sum_{i:\lambda_i\left(\frac{1}{n}\mathbf{A}\mathbf{A}^T\right)<\epsilon}\log\frac{\epsilon}{\lambda_i\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)} \\ &\leq \frac{\text{card}\left\{i : \lambda_i\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right) < \epsilon\right\}}{n}\log\frac{\epsilon}{\lambda_{\min}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T\right)}.\end{aligned}$$

This together with (57) and (56) implies that

$$\frac{1}{n}\log\det{}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)-\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)=O\left(\frac{\log^{2}n}{\sqrt{n}}\right) \tag{58}$$

with probability exceeding $1-\exp\left(-\hat{c}_{7}n\log n\right)$ for some constant $\hat{c}_{7}>0$

Besides, Equation (50) gives a lower bound on $\mathbb{E}\left(\frac{1}{n}\log\det^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\right)$. Setting $\epsilon=n^{-1/2}$ and $\delta=n^{-1/2}\log n$ in Lemma 2 leads to

$$\mathbb{P}\left(\frac{1}{n}\log\det{}^{\epsilon}\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)>(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+O\left(\frac{\log n+\log\epsilon}{n^{1/2}}\right)\right)\geq1-4\exp\left(-\tilde{c}n\log^{2}n\right). \tag{59}$$

Combining (58) and (59) with a union bound yields

$$\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^{T}\right)\geq(1-\alpha)\log\frac{1}{1-\alpha}-\alpha+O\left(\frac{\log^{2}n}{n^{1/2}}\right)$$

with probability exceeding $1-\exp\left(c_{9}n\log n\right)$ for some constant $c_{9}>0$.

Appendix G
Proof of Lemma 5

Suppose that the singular value decomposition of the real-valued $\boldsymbol{A}$ is given by $\boldsymbol{A}=\boldsymbol{U}_{\boldsymbol{A}}\left[\begin{array}{c}\Sigma_{\boldsymbol{A}}\\ \mathbf{0}\end{array}\right]\boldsymbol{V}_{\boldsymbol{A}}^{T}$, where $\boldsymbol{\Sigma}_{\boldsymbol{A}}$ is a diagonal matrix containing all k singular values of $\boldsymbol{A}$. One can then write

$$\begin{aligned}\log\det\left(\epsilon\boldsymbol{I}_{k}+\boldsymbol{A}^{T}\boldsymbol{B}^{-1}\boldsymbol{A}\right)&=\log\det\left(\epsilon\boldsymbol{I}_{k}+\boldsymbol{V}_{\boldsymbol{A}}\left[\begin{array}{cc}\Sigma_{\boldsymbol{A}} & \mathbf{0}\end{array}\right]\boldsymbol{U}_{\boldsymbol{A}}^{T}\boldsymbol{B}^{-1}\boldsymbol{U}_{\boldsymbol{A}}\left[\begin{array}{c}\Sigma_{\boldsymbol{A}}\\ \mathbf{0}\end{array}\right]\boldsymbol{V}_{\boldsymbol{A}}^{T}\right)\\&=\log\det\left(\epsilon\boldsymbol{I}_{k}+\Sigma_{\boldsymbol{A}}\left(\tilde{\boldsymbol{B}}^{-1}\right)_{[k]}\Sigma_{\boldsymbol{A}}\right)\geq\log\det\left(\frac{1}{n}\Sigma_{\boldsymbol{A}}^{2}\right)-\log\det\left\{\frac{1}{n}\left(\tilde{\boldsymbol{B}}^{-1}\right)_{[k]}\right\}\end{aligned} \tag{60}$$

where $\tilde{\boldsymbol{B}}=\boldsymbol{U}_{\boldsymbol{A}}^{T}\boldsymbol{B}\boldsymbol{U}_{\boldsymbol{A}}\sim\mathcal{W}_{m}\left(n-k,\boldsymbol{U}_{\boldsymbol{A}}^{T}\boldsymbol{U}_{\boldsymbol{A}}\right)=\mathcal{W}_{m}\left(n-k,\boldsymbol{I}_{m}\right)$ from the property of Wishart distribution. Here, $\left(\tilde{\boldsymbol{B}}^{-1}\right)_{[k]}$ denotes the leading $k\times k$ minor consisting of matrix elements of $\tilde{\boldsymbol{B}}^{-1}$ in rows and columns from 1 to k, which is independent of $\boldsymbol{A}$ from the Gaussianality property.

Note that $\frac{1}{n}\log\det\left(\frac{1}{n}\Sigma_{\mathbf{A}}^{2}\right)=\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}^{T}\boldsymbol{A}\right)$. Then Lemma 4 implies that with probability exceeding $1-\exp\left(-\tilde{c}_{9}n\log n\right)$, one has

$$\begin{aligned}\frac{1}{n}\log\det\left(\frac{1}{n}\Sigma_{\boldsymbol{A}}^{2}\right)&=\frac{1}{n}\log\det\left(\frac{1}{n}\boldsymbol{A}^{T}\boldsymbol{A}\right)=\frac{m}{n}\frac{1}{m}\log\det\left(\frac{1}{m}\boldsymbol{A}^{T}\boldsymbol{A}\right)+\frac{1}{n}\log\det\left(\frac{m}{n}\boldsymbol{I}_{k}\right)\\&\geq\alpha\left(-\left(1-\frac{\beta}{\alpha}\right)\log\left(1-\frac{\beta}{\alpha}\right)-\frac{\beta}{\alpha}\right)+\beta\log\alpha+O\left(\frac{\log^{2}n}{\sqrt{n}}\right)\\&=-\left(\alpha-\beta\right)\log\left(\frac{\alpha-\beta}{\alpha}\right)-\beta+\beta\log\alpha+O\left(\frac{\log^{2}n}{\sqrt{n}}\right)\\&=-\left(\alpha-\beta\right)\log\left(\alpha-\beta\right)-\beta+\alpha\log\alpha+O\left(\frac{\log^{2}n}{\sqrt{n}}\right).\end{aligned} \tag{61}$$

On the other hand, it is well known (e.g. [42, Theorem 2.3.3]) that for a Wishart matrix $\tilde{\boldsymbol{B}}\sim\mathcal{W}_{m}\left(n-k,\boldsymbol{I}_{m}\right)$, $\left(\tilde{\boldsymbol{B}}^{-1}\right)_{[k]}^{-1}$ also follows the Wishart distribution, that is, $\left(\tilde{\boldsymbol{B}}^{-1}\right)_{[k]}^{-1}\sim\mathcal{W}_{k}\left(n-m,\boldsymbol{I}_{k}\right)$. Applying Lemma 4 again

yields that

$$\begin{aligned}
&\frac{1}{n}\log\det\left(\frac{1}{n}\left(\tilde{\boldsymbol{B}}^{-1}\right)_{[k]}^{-1}\right)\\
=&\frac{n-m}{n}\frac{1}{n-m}\log\det\left(\frac{1}{n-m}\left(\tilde{\boldsymbol{B}}^{-1}\right)_{[k]}^{-1}\right)+\frac{1}{n}\log\det\left(\frac{n-m}{n}\boldsymbol{I}_k\right) &(62)\\
\leq&\left(1-\alpha\right)\left\{-\left(1-\frac{\beta}{1-\alpha}\right)\log\left(1-\frac{\beta}{1-\alpha}\right)-\frac{\beta}{1-\alpha}\right\}+\beta\log\left(1-\alpha\right)+O\left(\frac{\log n}{n}\right)\\
=&-\left(1-\alpha-\beta\right)\log\left(1-\frac{\beta}{1-\alpha}\right)-\beta+\beta\log\left(1-\alpha\right)+O\left(\frac{\log n}{n}\right) &(63)
\end{aligned}$$

holds with probability exceeding $1-\exp\left(-\tilde{c}_8 n\log n\right)$.

Combining (60), (61) and (60) leads to

$$\begin{aligned}
\frac{1}{n}\log\det\left(\epsilon\boldsymbol{I}_k+\boldsymbol{A}^T\boldsymbol{B}^{-1}\boldsymbol{A}\right)\geq&-\left(\alpha-\beta\right)\log\left(\alpha-\beta\right)+\alpha\log\alpha+\left(1-\alpha-\beta\right)\log\left(1-\frac{\beta}{1-\alpha}\right)\\
&-\beta\log\left(1-\alpha\right)+O\left(\frac{\log n}{\sqrt{n}}\right)
\end{aligned}$$

with probability exceeding $1-2\exp\left(-\tilde{c}_8 n\log n\right)$.

REFERENCES

[1] Y. Chen, Y. C. Eldar, and A. J. Goldsmith, "Channel capacity under sub-Nyquist nonuniform sampling," *under revision, IEEE Transactions on Information Theory*, April 2012. [Online]. Available: http://arxiv.org/abs/1204.6049

[2] P. L. Butzer and R. L. Stens, "Sampling theory for not necessarily band-limited functions: A historical overview," *SIAM Review*, vol. 34, no. 1, pp. 40–53, 1992.

[3] M. Mishali and Y. C. Eldar, "Sub-Nyquist sampling: Bridging theory and practice," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 98–124, Nov. 2011.

[4] H. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Mathematica*, vol. 117, pp. 37–52, 1967.

[5] C. Herley and P. W. Wong, "Minimum rate sampling and reconstruction of signals with arbitrary frequency support," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1555 –1564, Jul. 1999.

[6] R. Venkataramani and Y. Bresler, "Optimal sub-Nyquist nonuniform sampling and reconstruction for multiband signals," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2301 –2313, Oct. 2001.

[7] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[8] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, April 2006.

[9] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge University Press Cambridge, 2012, vol. 95.

[10] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 375 –391, Apr. 2010.

[11] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot, "Sparse sampling of signal innovations," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 31 –40, Mar. 2008.

[12] K. Gedalyahu, R. Tur, and Y. Eldar, "Multichannel sampling of pulse streams at the rate of innovation," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1491 –1504, April 2011.

[13] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. on Info Theory*, vol. 46, no. 3, pp. 933 –946, May. 2000.

[14] M. Medard and R. G. Gallager, "Bandwidth scaling for fading multipath channels," *IEEE Transactions on Information Theory*, vol. 48, no. 4, pp. 840 –852, Apr. 2002.

[15] P. Bello, "Characterization of randomly time-variant linear channels," *IEEE Transactions on Communications Systems*, vol. 11, no. 4, pp. 360 –393, Dec. 1963.

[16] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619 –2692, Oct 1998.

[17] G. D. Forney and G. Ungerboeck, "Modulation and coding for linear Gaussian channels," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2384 –2415, Oct 1998.

[18] S. Shamai, "Information rates by oversampling the sign of a bandlimited process," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1230 –1236, Jul. 1994.

[19] T. Koch and A. Lapidoth, "Increased capacity per unit-cost by oversampling," Sep. 2010. [Online]. Available: http://arxiv.org/abs/1008.5393

[20] M. Peleg and S. Shamai, "On sparse sensing of coded signals at sub-Landau sampling rates," *IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, pp. 1 –5, Nov. 2012.

[21] Y. Chen, Y. C. Eldar, and A. J. Goldsmith, "Shannon meets Nyquist: Capacity of sampled Gaussian channels," *accepted to IEEE Transactions on Information Theory*, 2013. [Online]. Available: http://arxiv.org/abs/1109.5415

[22] A. Papoulis, "Generalized sampling expansion," *IEEE Transactions on Circuits and Systems*, vol. 24, no. 11, pp. 652 – 654, Nov 1977.

[23] A. E. Gamal and Y. H. Kim, *Network Information theory*. Cambridge University Press, 2011.

[24] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels," *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1229–1241, 1959.

[25] M. Effros, A. Goldsmith, and Y. Liang, "Generalizing capacity: New definitions and capacity theorems for composite channels," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3069 –3087, July 2010.

[26] D. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," Dec. 2011. [Online]. Available: http://arxiv.org/abs/1112.0708

[27] Y. Wu and S. Verdu, "Optimal phase transitions in compressed sensing," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6241 –6263, Oct. 2012.

[28] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed Sensing, Theory and Applications*, pp. 210 – 268, 2012.

[29] A. Guionnet and O. Zeitouni, "Concentration of the spectral measure for large matrices," *Electronic Communications in Probability*, no. 5, pp. 119 –136, 2000.

[30] A. Ghosh, J. Zhang, R. Muhamed, and J. G. Andrews, *Fundamentals of LTE*. Pearson Education, 2010.

[31] "LTE in a nutshell: White paper on physical layer," 2010.

[32] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer, 1998.

[33] P. Massart and J. Picard, *Concentration Inequalities and Model Selection*, ser. Lecture Notes in Mathematics. Springer, 2007.

[34] T. Tao and V. Vu, "Random covariance matrices: Universality of local statistics of eigenvalues," *The Annals of Probability*, vol. 40, no. 3, pp. 1285 – 1315, 2012.

[35] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[36] T. Tao, *Topics in Random Matrix Theory*, ser. Graduate Studies in Mathematics. Providence, Rhode Island: American Mathematical Society, 2012.

[37] M. Mishali and Y. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 993 –1009, March 2009.

[38] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 1985.

[39] M. Aoki, *New Approaches to Macroeconomic Modeling*. Cambridge University Press, 1998.

[40] A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, "Smallest singular value of random matrices and geometry of random polytopes," *Advances in Mathematics*, vol. 195, no. 2, pp. 491 – 523, 2005.

[41] M. Rudelson and R. Vershynin, "Smallest singular value of a random rectangular matrix," *Communications on Pure and Applied Mathematics*, vol. 62, no. 12, pp. 1707–1739, 2009.

[42] Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu, *Multivariate Statistics: High-dimensional and Large-sample Approximations*, ser. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley, 2010.

[43] Z. Chen and J. J. Dongarra, "Condition numbers of Gaussian random matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 3, pp. 603–620, July 2005.